Integrating Digital Humanities into the Web of Scholarship with SHARE

November 21, 2019

Cynthia Hudson-Vitale, Penn State University Libraries

Judy Ruttenberg, Association of Research Libraries

Matthew Harp, Arizona State University Library

Rick Johnson, Hesburgh Libraries, University of Notre Dame

Joanne Paterson, Western Libraries, Western University

Jeffrey Spies, 221B

Table of Contents

NATIONAL ENDOWMENT FOR THE HUMANITIES

This project has been made possible in part by a generous grant from the National Endowment for the Humanities: Exploring the human endeavor.

Executive Summary

The project Integrating Digital Humanities into the Web of Scholarship with SHARE (2017–2019) was designed to investigate the value SHARE could have for digital humanities (DH) scholars, by exploring how scholars promote discovery of their own DH work, and how they find digital scholarship or its components for their own use. The project leaders' assumptions were that (1) discovery of DH scholarship was difficult because it relied on web discovery through keywords rather than structured metadata, and (2) structured metadata and improved discovery were essential for enabling the enduring stewardship of DH scholarship by the research library community.

The project involved a mixed-methods approach and four phases of work:

- An online survey of the digital humanist's workflow
- A design workshop of DH practitioners and librarians to create wireframes of potential discovery solutions
- A series of focus groups at DH centers and libraries to explore attitudes and practices
- Creation of prototypes to test some of the workshop-generated designs

Key Findings

Survey of Digital Humanist's Workflow

An online survey captured the full workflows of 42 different DH projects. Workflow questions used the TaDiRAH research activity terms: Capture, Creation, Enrichment, Analysis, Interpretation, Storage, Dissemination, and Meta-Activities.[1] A full report and analysis of the survey is in **Appendix B**. Key findings from the survey include the following:

- More than 100 different tools and software were used across the identified projects. The greatest variety of tools and software was

found in the Create/Capture phase of the DH workflow, where 50 distinct tools were in use. Activities under these headings include imaging, recording, writing, translating, programming, web development, and more.
- More than 71% of survey respondents indicated their willingness to share their DH project assets in some way. Respondents also indicated they were most likely to share the raw assets that were captured through the digitization process or created through the project period. The majority of respondents shared their assets on GitHub and on personal websites.
- Some respondents reported on projects in fields not usually considered to be humanities, such as sociology and anthropology. While additional research is necessary, this supports the idea that digital humanities scholarship or methods may be discipline agnostic, and not tied to the traditional divisions of research in academia.
- Respondents were asked to rank 12 potential search filters on a Likert scale of "not at all important" to "extremely important." The most important filter, or search criteria, was whether the project was open access.

Workshop

A two-day workshop in February 2018 included panel discussions, lightning rounds, small-group work, and active design. Among the findings from the workshop were the following:

- Overdependence on digital scholarship project websites presents both long-term storage/archiving challenges and technical challenges.
- Citation and reuse are important to scholars, and there is no easy way to communicate use of a web project in teaching or research (no citation registry).
- The licensing and reuse environment is complex and not well-disclosed or understood for digital humanities projects.

- Knowing the provenance of digital scholarship is essential for its reuse in a scholarly context, and this is often difficult or opaque.
- Scholars and librarians need markers of completion that enable digital projects to become publications that enter the library's stewardship workflow.

Focus Groups

The project team visited six institutions in May and June 2018. A full report and analysis of the focus groups is in **Appendix D**. Highlights from the focus groups include the following:

- There are not clearly delineated points where digital projects move from a center or lab to the library.
- Across focus groups, there was consensus that ambiguity exists around which digital projects constitute "data" and which "collections." The two types of projects are treated differently for purposes of discovery and long-term stewardship. Many focus group participants called for development of community guidelines to develop this distinction and its criteria.
- Support for digital humanities projects is often handled in an ad hoc manner and based on existing skills, scope, and financial resources.

Introduction

SHARE began in 2013 as part of a growing open academic infrastructure movement. Like many community software projects, it was funded by a combination of federal and private grants, and organized by a collaborative group of associations, universities, and libraries. SHARE's open source software harvests and indexes free, open metadata about scholarship and links scholarly activity across the research life cycle and in any discipline. Project leaders proposed that by working closely with scholars and librarians in the DH community, SHARE could offer a means for scholars to interlink all the components of their work; for librarians to have a means to accurately track usage of all the components of a DH project; for scholars and students to quickly find the relevant scholarship and primary sources they need; and for new project leaders to quickly gain an understanding of all the existing content and tools at their disposal.

This project demonstrated that those who work in digital humanities would embrace improvements to help them discover tools and data to use in new projects, but that today the complexity of formats, metadata structures, and discovery channels makes that difficult. In addition, the actual degree of demand for content and data is far from clear. The progress made so far appears to be taking place where there are disciplinary-driven needs—scholars in a field who have a clear need and desire to be aware of other work that is relevant to them. Approaching discovery from a disciplinary point of view appears to run the risk of further siloing information we are seeking to share and integrate more broadly. That said, as a starting point, having scholars work collaboratively, perhaps with adjacent fields, may offer the best chance to establish crosswalks among subjects, a pathway to fully SHARED data.

Project Activities

Survey: Digital Humanist's Workflow

In the fall of 2017 the project team undertook the first step toward better understanding DH workflow habits by developing and administering a survey. Questions for the survey were compiled by project team members and the advisory board in October 2017. Workflow questions used research activities terms from the Taxonomy of Digital Research in the Humanities (TaDiRAH):

> This taxonomy of digital research activities in the humanities has been developed for use by community-driven sites and projects that aim to structure information relevant to digital humanities and make it more easily discoverable. The taxonomy is expected to be particularly useful to endeavors aiming to collect information on digital humanities tools, methods, projects, or readings.[2]

TaDiRAH contains more than 121 terms and definitions that are used to give DH scholarship a common language. For the purposes of this survey the project team used the following activities, in some cases combining activities, to describe DH scholarship:

- Capture/Creation
- Enrich
- Analyze
- Interpret
- Store/Disseminate

For a full analysis and report, see **Appendix B**.

Workshop

In mid-February 2018 during the American Library Association (ALA) Midwinter Meeting in Denver, Colorado, the project team convened approximately 30 invited digital humanists, librarians, publishers, and managers of disciplinary and institutional repositories and related, large-scale, digital library projects to work on challenges in discovering

DH scholarship. While some of those challenges are well-known—for example, the tendency for DH projects to live on stand-alone websites rather than in scholarly repositories, and the related complications of identifying and describing component parts of the work—the interventions and services to address them within the scholars' workflow are less documented and less well-understood. Through a combination of lightning talks and active working sessions, participants collectively defined key issues around DH project metadata, and how SHARE's metadata-harvesting technology might integrate with the world of DH registries, identifiers, and repositories to improve discovery.

The workshop was facilitated by Nancy Maron of BlueSky to BluePrint, and the participants' collective work informed the next phases of the project, including site visits by project team members to DH centers (either based in or outside of libraries) and prototyping discovery tools using the SHARE harvester with institutional and/or disciplinary repositories or registries.

The group worked on a series of exercises to (1) define the problem of distributed DH assets, (2) envision ideal interfaces for DH discovery, and (3) consider the underlying data structures necessary to realize those interfaces. The workshop agenda and participant materials are in **Appendix C**.

Ideas generated during the workshop were recorded for further exploration during the campus site visits. The team solicited volunteers and suggestions from the workshop participants and subsequently asked the advisory board to help rank them. One participant noted that such research-based site visits are valuable to DH advocates on campus, enabling the case for further investment in DH and potentially in a repository.

The workshop opened with an introduction to SHARE, and to the project and its basic proposition:

> Contemporary scholarship is interdisciplinary, multimodal and distributed across a wide network of tools, repositories, and websites. A digital humanities (DH) project may produce more than one manuscript

(books or articles), each published on a different publisher's website, grant award information, any number of preprints on MLA CORE or other services, data sets and code books on Dryad or Figshare, and text mining or cleaning scripts on GitHub. By linking these dispersed research objects with one another, they can be evaluated and understood as part of the same intellectual work, thus increasing our understanding of the scholarship and limiting any intellectual stratification in the community of scholars based on their different contributions (collections of primary sources, computational research, and tools, e.g.). On the other hand, when individual project components such as scripts are too tightly bundled and isolated on project websites, they are hidden from networked search and discovery tools, resulting in similar problems.

The design question for the group was: How can we enhance discoverability for DH work by looking at metadata creation, generation, and capture?

A panel of experts helped define the problem:

- **Nikolaus Wasmoen**, Visiting Assistant Professor in Digital Humanities, University at Buffalo
- **Quinn Dombrowski**, Digital Humanities Coordinator, University of California, Berkeley
- **Annie Johnson**, Library Publishing and Scholarly Communications Specialist, Temple University

Panelists identified the following challenges for DH projects:

- Web hosting and file storage
- "Most faculty don't see their work as data"
- Desire to search across platforms and projects
- Lack of resources for database design
- Lack of resources for digital repositories, particularly institutional repositories

Asked what they **wish** they could know about DH projects on the web, panelists and participants offered:

- What is the extent of peer review that the project has undergone and at what stage?
- How can people reuse component parts outside the original use case?
- What format are the files in?
- What projects are appropriate for teaching, or even designed with pedagogy in mind?
- What are the reuse rights and/or permissions, including cultural restrictions?
- What else did the project leaders work on? Who influenced them? ("I want a profile of the researcher.")
- What is the provenance of the digitized materials?
- How accessible to people with print disabilities are the materials and projects?

Participants worked in small groups to develop user stories based on the challenges and desires raised in the morning session. Instructions for groups were to create user stories with the definition of the type of user and their needs, and then asked to consider, "wouldn't it be cool if?"

Sample user stories:

1. **Type of user**: Graduate students and faculty

   User need: Digitization of library materials, models and examples, technical and design consultation for project

   Wouldn't it be cool if there were a well-organized metadata format that is interoperable, tools for using it with common platforms like Omeka, and a practice of making data management plans?

2. **Type of user**: Early career

   User need: Credentials, improvement in scholarly profile

   Wouldn't it be cool if their project appeared in a registry, with a digital object identifier (DOI) and a badge?

3. **Type of user**: Teacher/professor

   User need: Open, reusable materials, sample activities

   Wouldn't it be cool if digital humanities projects were tagged as teaching materials?

4. **Type of user**: Academic administrator

   User need: Evidence to demonstrate the impact of scholarship, enhance institutional reputation, benchmark against peer institutions

   Wouldn't it be cool if there was a dashboard that aggregated citations and uses of projects, with institutional identifiers?

Small groups continued their work by turning their user stories into wireframes by designing solutions or steps to improve discovery. Ideas included the following:

- **I Used This!**: a button to automatically convey reuse to a project creator
- **Omeka S plugin**: to capture and expose a core set of bibliographic project metadata
- **Teaching dashboard**: to find DH teaching materials and activities
- **Researcher profile**: to provide context on research projects by linking the researchers to their other work
- **Impact dashboard**: to track reuse of DH project by person and institution

The workshop concluded with collective reflections, recommendations, and considerations for the community and the project team:

- Much DH support work, rather than being seen as bespoke, can be understood as core to established library workflows in research data management, web archiving, and publishing (including presses, next-generation repositories, or other).
- Librarians should work within publishing and cataloging structures to develop ways to capture key metadata fields.

- The community needs guidelines for ongoing stewardship of DH work that does not end up in a "published" DH project.
- The community should draw upon existing codes of best practice, such as for fair use, visual materials, and software preservation, and continue to develop additional resources.

Focus Groups

In May–June 2018, the project team visited six US and Canadian universities and held focus groups comprised of DH librarians, scholars, and DH center staff. The project team selected the six campuses in consultation with the project advisory board and with the 30 participants of project workshop, held that February. Criteria included (1) presence and location of the DH center (in or outside the library), (2) geographic diversity, (3) presence of a Council on Library and Information Resources (CLIR) Postdoctoral Fellow, (4) reputation for innovation in DH, and (5) inclusion of institutions that are less frequently profiled. In addition to facilitating the focus groups, project leaders emailed contacts at the six institutions, supplied email text to invite participants, a registration form, and offered a small budget for catering. For a full report and analysis of the focus groups, see **Appendix D**.

Prototypes

This project funded the development of two prototypes related to DH discovery and metadata, one by Penn State University Libraries and the other by 221B.

Penn State University Libraries

This prototype tested two products to scrape published websites for metadata and map the terminology used in the test sites, the Dublin Core metadata used by the Omeka platform, and the SHARE metadata schema.

Project deliverables: "SHARE-ing Omeka in the Web of Digital Scholarship," by Michael Roth, 2018, https://osf.io/yfp39/; and

"Webscraper.io: A How-to Guide for Scraping DH Projects," by Michael Roth, 2018, https://osf.io/dpk5w/.

Webinar: "SHARE-ing Omeka in the Web of Digital Scholarship," October 4, 2018, 56:21, http://www.share-research.org/2018/10/webinar-recording-share-ing-omeka-in-the-web-of-digital-scholarship/
Michael Roth, Project Intern
Heather Froehlich, Literary Informatics Librarian
Cynthia Hudson-Vitale, Head of Digital Scholarship and Data Services

221B Consulting Group

This prototype is a dashboard of National Endowment for the Humanities (NEH) digital humanities awards, enabling a robust search of awards, for example by program area, institution, gender of project lead, and to explore resolvable URLs of project websites and languages used on GitHub product URLs.

Project deliverable: https://github.com/221B-io/neh-dashboard/

The prototype team harvested NEH past award data. Making specific use of grant products relationships, for each grant with grant products, the team collected information about the type of products and its URL. This information was used to describe where grant products were being archived—using the term loosely. Each URL was then resolved, to check whether or not the product was still accessible. For products linking to GitHub, the team used the GitHub API to find out if the product was still being developed after the grant ended and what programming languages were being used across NEH grants. Along with grant products, the team created visualizations showing funding by institution as well as funding by gender of principal investigator over time, using name to infer gender. Finally, the team used Elasticsearch to index the data harvested in order to create an interactive dashboard for exploring NEH funding.

Scripts available on GitHub can be used to do the following:

- Request data from the NEH website
- Parse the HTML files saved locally into JSON files
- Add the gathered products into a new JSON file with their respective grants, along with some computed indices and URL resolution data (This is what will be uploaded to Elasticsearch.)
- Turn the JSON resulting from the above command into a properly formatted bulk-request body for Elasticsearch

Presentation: "Prototypes for Enhancing the Discoverability of Digital Humanities Scholarship," by Judy Ruttenberg, Cynthia Hudson-Vitale, and Jeffrey Spies, at the Coalition for Networked Information (CNI) Fall 2018 Membership Meeting in Washington, DC.

Conclusion

There were two principal lines of inquiry in this project: (1) scholar and librarian requirements for improving discovery of DH projects; and (2) whether SHARE could be part of an improved discovery environment. With respect to the latter, while this project informed the future development of SHARE, the Association of Research Libraries (ARL) stepped back from its role as SHARE's product owner in 2018 and the development is now led by community members in ARL institutions. Consequently, ARL is no longer directing resources for SHARE's development. However, the NEH grant dashboard prototype developed for this project contributed to SHARE's codebase and set of tools for a university dashboard that uses Elasticsearch for indexed, harvested linked data for visualizations and analysis.

With respect to questions of metadata and discovery, across focus groups, there was consensus that ambiguity exists around which digital projects constitute "data" and which are "collections." The two types of projects are treated differently for purposes of discovery and long-term stewardship. Many focus group participants called for development of community guidelines to develop this distinction and its criteria. ARL, through its Scholars and Scholarship priority area, is committed to working with disciplinary communities, such as scholarly and learned societies, to craft such guidelines.

People can easily describe "problems to solve," particularly when it comes to the topic of managing digital humanities projects. Anyone who has led, worked on, funded, or managed someone who runs a DH project, knows that there are many challenges involved in their building, maintenance, growth, and preservation.

This planning grant sought to pinpoint one set of challenges—discovery and persistent stewardship—that are a challenge for content of any sort, but seem to be exaggerated for DH projects, which can involve multiple formats, have many collaborators but no one owner, ill-defined "outcomes," and uncertain means of assessing success.

Through an online survey, and by talking with dozens of people through an in-person workshop and six university-based focus groups, we hoped to develop a keener sense of where the demand for better discovery and preservation might be greatest, and to generate some ideas about how SHARE might offer the potential for some solutions.

What We Learned

By speaking with many people who have built and managed DH projects, the project team learned a great deal about the range of challenges faced, as well as what participants in the project felt would need to be improved. Suggested improvements included the following:

1. A clearer sense is needed, among both DH practitioners and managers, of what "discovery" actually means. Some defined this as a proactive act of outreach (promotion); others felt this was something that shared metadata schemas and practices would resolve. Both are extremely important though SHARE is best placed to support efforts to improve the latter.

2. Specifically concerning metadata, participants emphasized the challenges in having so many varied taxonomies, ontologies across the full range of topics and disciplines. At least two of the focus groups included discussion of various communities of practice that have begun to emerge as a means of providing coordinated access—sometimes through agreed upon discipline-specific taxonomies—to the many types of scholarly outputs DH encompasses. Support for strengthening communities of practice, so that metadata suits the topics at hand, is an area SHARE is well placed to support.

3. Whether the "catalog" is SHARE, WorldCat, or something else, participants expressed the need to have something directory-like, as a means of identifying what those DH works were. Some pointed to the Directory of Open Access Journals (DOAJ) and the Directory of Open Access Books (DOAB) as models to examine. As one participant pointed out, it is important that whatever solution is chosen needs to have wide appeal and be widely known, for this to work.

4. Today, much of the work of DH discovery, according to those we spoke with, appears to rely upon a network of relationships. As one participant pointed out, “You just have to know your people, your sources. You have to know who works with whom.”

5. Assuming there were to be agreement on a common metadata standard, or standards; and agreement on what entity would be serving the central organizing function, one participant raised the question of who ought to generate the metadata. This issue of who is best placed to do this work, what time it requires, and what the vetting process (if any) would be, are questions worth further exploration.

6. What would people want to be able to find? While not all groups addressed this in detail, some of the things that surfaced included: names, dates, places. Others spoke about criteria for evaluation of DH works (assuming this would mean a peer review process prior to listing?). Some criteria might include: “project has an identifiable manager with contact information, and an editorial board or editorial team, or league with scholarly credentials. It’s clearly identified as original or taken from other sources like archives. Provides full citation information...includes a bibliography...says when it was last updated...has a long-term preservation plan in place.”

7. Finally, while a great deal of time was spent talking about challenges, the discussion concerning demand for the materials-to-be-discovered was sobering. According to one participant, “I don’t think data is heavily reused in the humanities at this point. The tools are reused, so we use a lot of open tools. And those tools do get reasonable circulation...But actually taking data sets and managing them with new data sets and sharing them and enriching them and stuff like that, I don’t see a heck of a lot of that kind of reuse, not so much in libraries, either. We tend to produce metadata and push it out to other central locations. We don’t ingest a lot of other people’s data, really, even in the libraries.”

This project clearly demonstrated that those who work in digital humanities would embrace improvements to help them discover tools and data to use in new projects, but that today the complexity of formats, metadata structures, and discovery channels makes that difficult. In addition, the actual demand for content and data is far from clear. The progress made so far appears to be taking place where there are disciplinary-driven needs—scholars in a field who have a clear need and desire to be aware of other work that is relevant to them.

Approaching discovery from a disciplinary point of view appears to run the risk of further siloing information we are seeking to share more broadly! That said, as a starting point, having scholars work collaboratively, perhaps with adjacent fields, may offer the best chance to establish crosswalks among subjects, a pathway to fully SHARED data.

Endnotes

1. TaDiRAH—Taxonomy of Digital Research Activities in the Humanities, accessed November 14, 2019, http://tadirah.dariah.eu/vocab/index.php.
2. "About...," TaDiRAH—Taxonomy of Digital Research Activities in the Humanities, accessed November 14, 2019, http://tadirah.dariah.eu/vocab/sobre.php.

Appendix A: Project Advisory Board

Nancy L. Maron, BlueSky to BluePrint
Natsuko Nicholls, Institute for Research on Innovation and Science
Thomas Padilla, University of Nevada Las Vegas Libraries
Christa Williford, Council on Library and Information Resources

Appendix B: Survey Report

The Digital Humanist's Workflow: Survey Results

This report presents the outcomes and results of "The Digital Humanist's Workflow" survey completed as part of the "Integrating Digital Humanities into the Web of Scholarship with SHARE" project funded by the National Endowment for the Humanities.

Authors:

Cynthia Hudson-Vitale, Penn State University Libraries
Judy Ruttenberg, Association of Research Libraries
Matthew Harp, University of Arizona Library
Joanne Paterson, Western Libraries, Western University
Richard Johnson, Hesburgh Libraries, University of Notre Dame
Jeffrey Spies, 221B

Advisory Board:

Nancy Maron, Bluesky to Blueprint
Natsuko Nicholls, Institute for Research on Innovation and Science
Thomas Padilla, University of Nevada Las Vegas Libraries
Christa Williford, Council on Library and Information Resources

Background

Digital humanities (DH) scholarship and research produces a number of digital assets that can be stored and shared in distributed locations. Much like researchers in the social sciences and the STEM fields, digital humanists have distinct digital outputs such as code, analytical scripts, tabular data, images, and more. They also produce related resources such as grant information, publications, and digital product or data management plans. All of these outputs and resources may be shared on different platforms online, which unfortunately limits their discoverability. For DH in particular, this is further complicated by the fact that digital humanists typically share their DH projects on stand-alone websites, which are not included in typical scholarly discovery indexes. However, DH projects should be findable, accessible, interoperable, and reusable (FAIR).

Addressing the challenge of enhancing the FAIR-ness of DH scholarship is not a new endeavor. Maron and Pickle (2014) analyzed the landscape of DH project production on campus from an institutional perspective—looking at service and research models, level of cross-unit coordination or lack thereof, the kinds of assets being produced, and faculty priorities for their stewardship over the long term. Preservation of digital projects and their components was an abiding concern. Project directors described themselves as dependent upon outreach efforts to maintain visibility and therefore garner ongoing technical and content support for their projects.

SHARE—built as a partnership between the Association of Research Libraries (ARL) and the Center for Open Science (COS) and now a community-driven project—is part of a growing open academic infrastructure movement. SHARE's open source software aggregates

free, open metadata about scholarship and links scholarly activity across the research life cycle and in any discipline. By working closely with scholars and librarians in the DH community, SHARE can offer a means for scholars to link all the components of their work; for librarians to accurately track usage of all the components of a DH project; for scholars and students to quickly find the relevant scholarship and primary sources they need; and for new project leaders to gain an understanding of all the existing content and tools at their disposal. Aggregating information about DH projects and components will also assist campus units in optimizing their resources and services in support of this scholarship.

As a first step in addressing this need, the project team undertook a survey of digital humanist researchers, content creators, and librarians. This survey sought to address the following goals:

1. Identify common tools used by researchers across the DH landscape that are used throughout the DH project life cycle as defined by the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH)
2. Identify where DH project assets are typically shared
3. Identify DH researcher search and discovery habits

Methods

If we are to increase the FAIR-ness of DH scholarship, then we need to investigate workflows, search habits, challenges, and potential solutions. To better understand the extent of this issue, the planning grant adopted a mixed-methods approach comprised of a survey, a workshop, and focus groups. The outcomes of this planning will include a white paper on recommendations to enhance DH FAIR-ness and a number of prototypes.

In the fall of 2017 the project team undertook the first step to better understand DH workflow habits by developing and administering a survey. Project team members and the advisory board compiled questions for the survey in October 2017. Workflow questions used the

TaDiRAH research activities terms. From the TaDiRAH website:

> *This taxonomy of digital research activities in the humanities has been developed for use by community-driven sites and projects that aim to structure information relevant to digital humanities and make it more easily discoverable. The taxonomy is expected to be particularly useful to endeavors aiming to collect information on digital humanities tools, methods, projects, or readings.*

TaDiRAH has over 121 terms and definitions that it uses to give DH scholarship a common language. For the purposes of this survey the project team used the following activities, in some cases combining activities, to describe DH scholarship:

- Capture/Create
- Enrich
- Analyze
- Interpret
- Store/Disseminate

The project team also developed questions that sought to uncover the prevalence of data sharing, what kinds of assets were typically shared, and where those products were shared. From a workflow perspective, this information assisted the project team members in better understanding the extent to which DH scholarship is distributed and to where. The project team was additionally interested in alternative forms of searching, such as by peer review status, data creation method, tools used, year of creation, temporal span, etc. Prior to releasing the survey, the project advisory board and DH researchers at Washington University in St. Louis reviewed it.

The survey ran from October 26, 2017, to November 30, 2017, and again in preparation for a DH workshop held in February 2018. We distributed it on DH listservs and Slack groups and widely circulated it on Twitter and library email lists. We used the Qualtrics survey platform to host and distribute the survey.

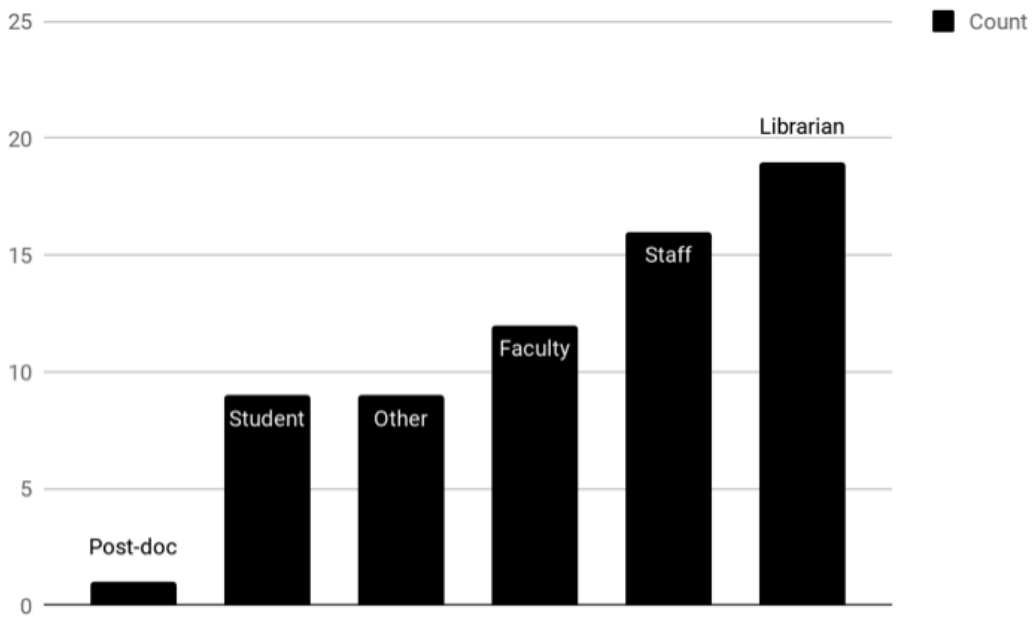
The survey comprised approximately 25 questions. The survey instrument may be found online at: https://osf.io/kx2cy/.

Results

In total, we collected 93 responses. From these we were able to map the full workflows of 42 different DH projects.

Survey Response Context

The projects came from a variety of content creators, including faculty, staff, and librarians. The "Other" field included responses such as Director, Head, Consultant, and Specialist, to name a few.

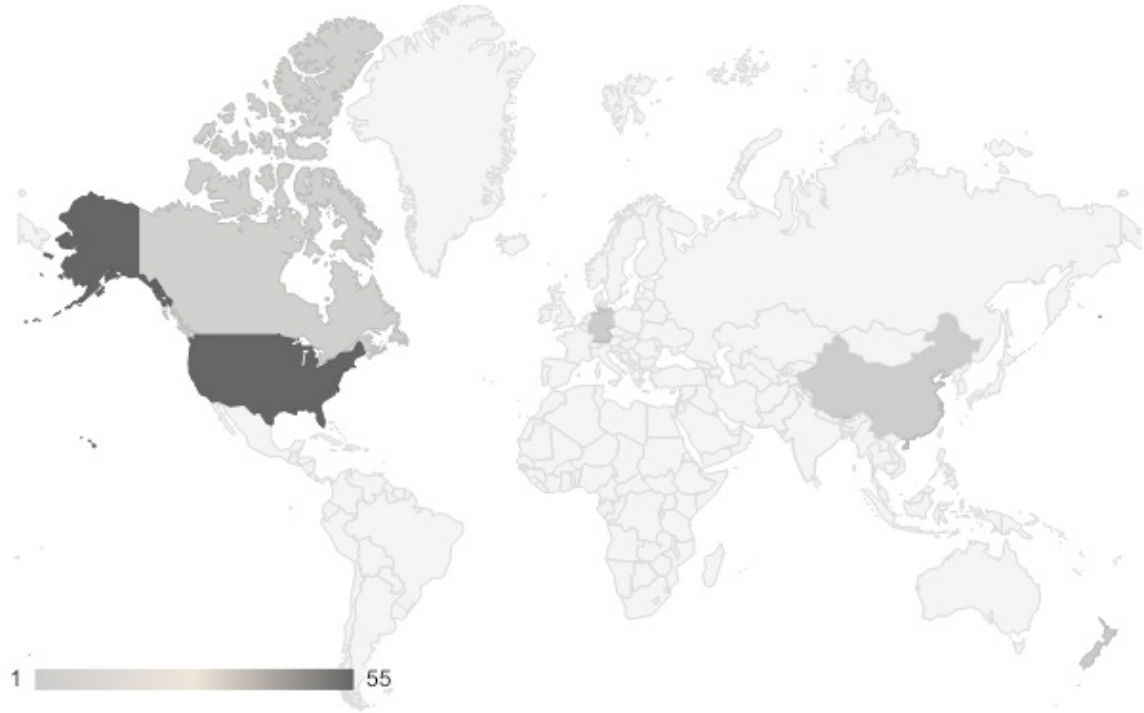


The distribution of disciplines covered by the survey was fairly varied across the humanities and into the social sciences of information and library sciences. In total, there were over 40 different disciplines in the survey, with many responses indicating work that covered more than one discipline. The top ten disciplines are listed below.

Discipline	Response Count
History	23
Digital Humanities	10

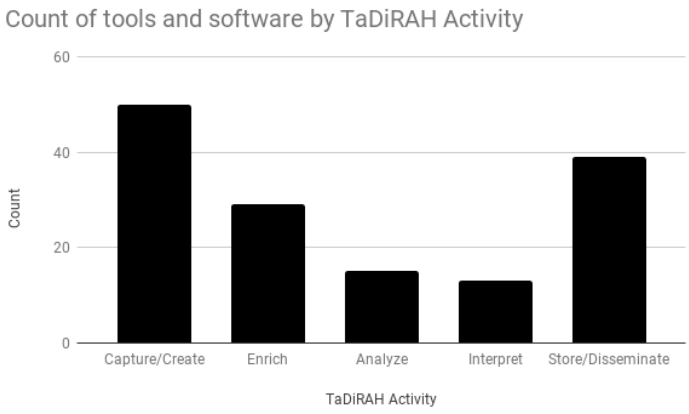
English	12
Library Science	12
Humanities	10
Art History	5
Classics	5
Sociology	4
Anthropology	3
Philosophy	3

There was also a strong geographic distribution in the responses. Eighteen were from the Northeast, eleven from the Midwest, seven from the Western United States, and nineteen from the Southern United States. In addition, eleven individuals from international organizations completed the survey, including digital humanists from Great Britain, Canada, New Zealand, China, and Germany. The type of institution responding also varied, with responses coming from small liberal arts colleges, nonprofit organizations, large research institutions, and more.



Workflows

Results from the workflow data indicate that digital humanists use a wide variety of tools and software throughout a DH project, independent of where it falls along the DH life cycle. We identified over 100 different tools and software across the identified DH projects. At a high level, we see tool usage breakdown as follows:

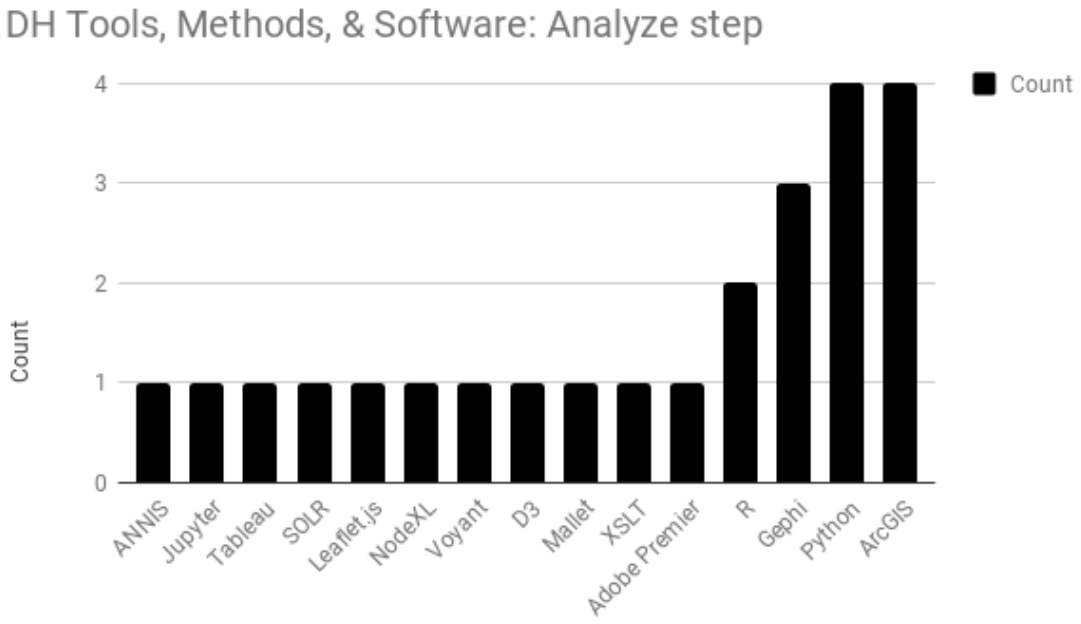


We found the greatest variety of tool and software use in the Capture/Create phase of the DH workflow, where respondents named 50 distinct tools. As you can see from the visualization, Store/Disseminate [n=39] and Enrich [n=29] had the second and third greatest tool usage.

Capturing and creating DH scholarship can mean the activity of capturing digital surrogates of existing cultural artifacts or producing born-digital assets. Survey results indicated the Capture/Create activity included such tools, methods, and software as Python [n=11], oXygen [n=6], and ABBYY FineReader [n=5]. Less frequently used approaches included R, Zotero, Audacity, and Microsoft Access. One-off approaches to capture and create assets related to DH projects included ArcGIS, Handbrake, Dragon, and FileZilla.

As defined by TaDIRAH, enrichment "refers to the activity of adding information to an object of enquiry, by making its origin, nature, structure, meaning, or elements explicit. This activity typically follows the capture of the object." (http://tadirah.dariah.eu/vocab/index.php?tema=21&/3_enrichment). Enriching digital assets to support DH research also involved a number of different tools and software. The survey found that the most heavily used tool to enrich DH scholarship was OpenRefine [n=10], followed by Python [n=7], R [n=3], XSLT [n=3], and oXygen [n=3].

The analyze step involves extracting information or meaning from a digital collection. Results from this step show that the most frequently used tools, methods, or software were Python n=[4] and ArcGIS [n=4], followed closely by Gephi [n=3] and R [n=2]. There were also eleven tools, methods, or software that were mentioned only once, including Tableau, D3, and Leaflet.

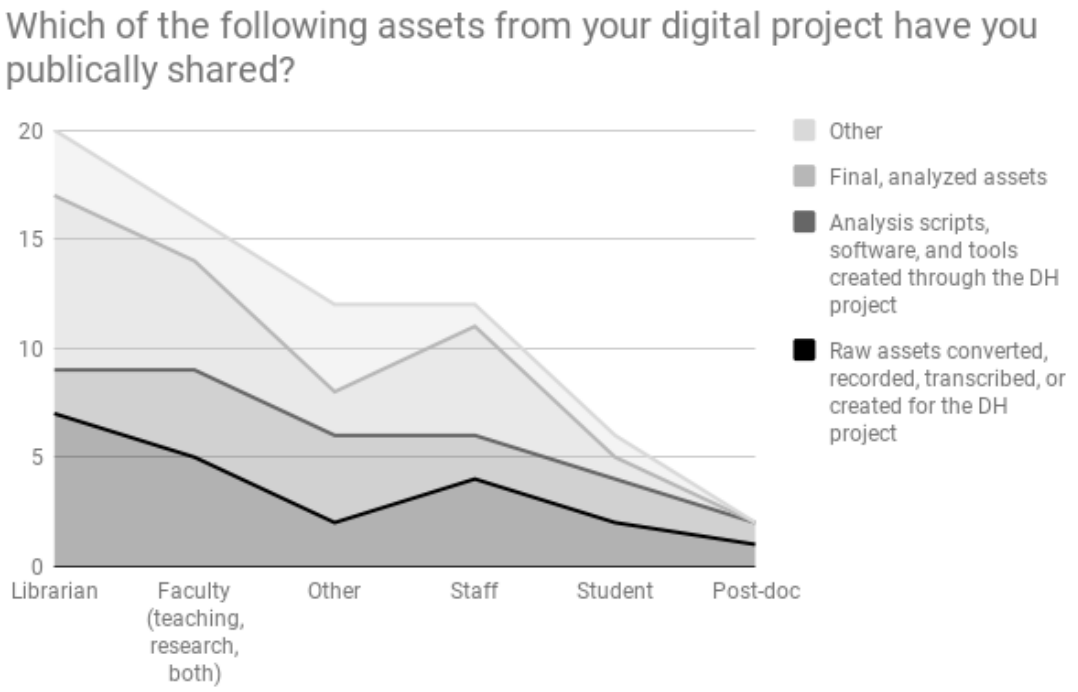


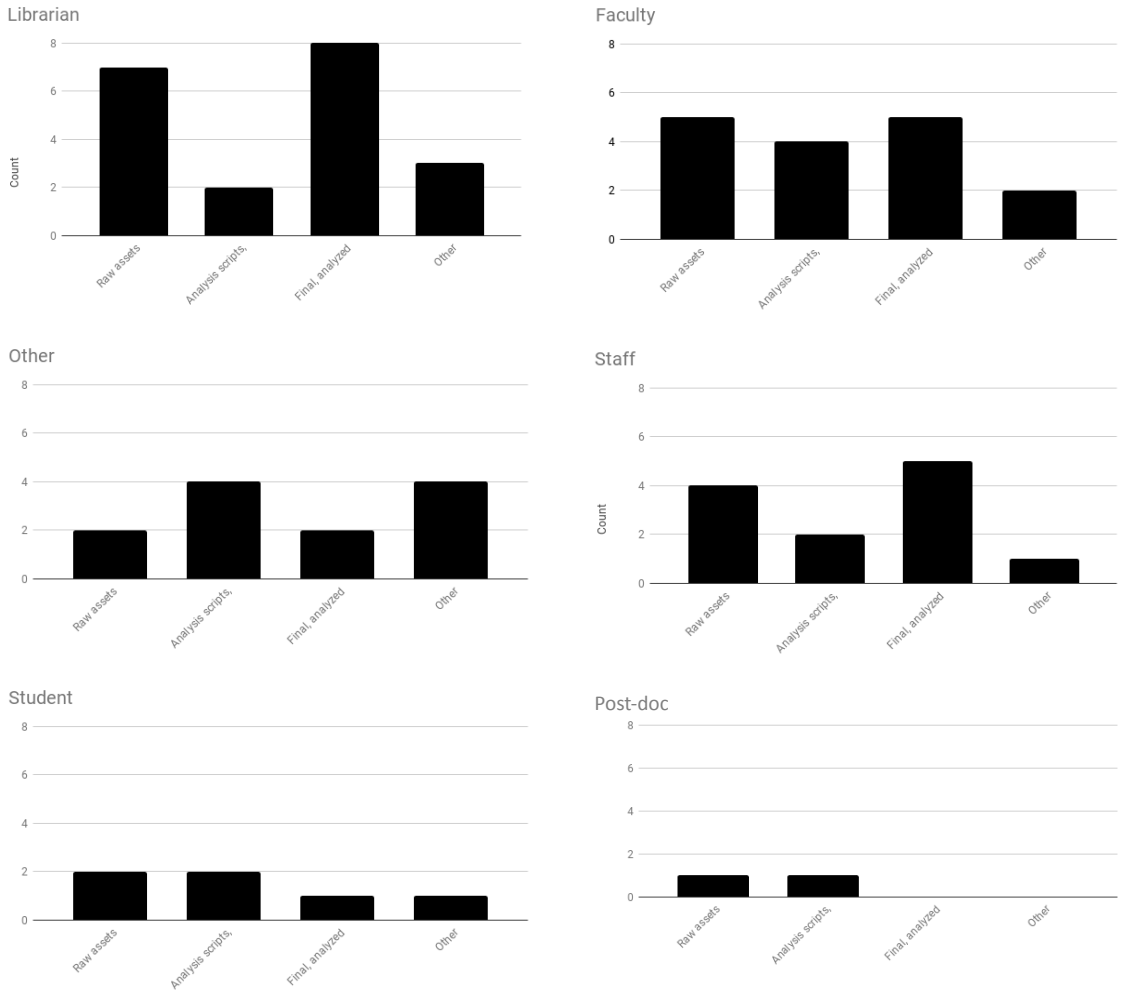
TaDiRAH defines interpretation as "the activity of ascribing meaning to phenomena observed in analysis." (http://tadirah.dariah.eu/vocab/index.php?tema=33&/5_interpretation). The Digital Humanists Workflow survey results indicated that respondents used 12 different tools, methods, or software across the 42 different DH projects reported on through this survey. They used R [n=4] the most heavily for interpreting DH projects, followed closely by D3 [n=2] and Python [n=2].

The final stage of the DH scholarship workflow that we surveyed was storage and dissemination, which involves making digital copies of the scholarship accessible to the project team and/or the public. Breaking down the store and disseminate activity, we found 39 unique tools in use for the DH projects. The tools that respondents most heavily used

for the storage and dissemination of DH scholarship were GitHub and Ruby, with eight mentions each, with websites and WordPress also mentioned frequently. Respondents listed over 26 different tools, software, and methods as being used only once across all projects analyzed.

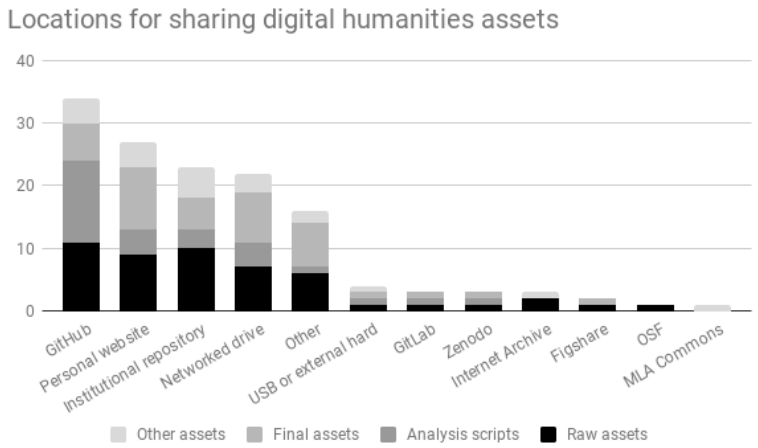
Over 71% of respondents indicated their willingness to share DH project assets. Respondents also indicated they were most likely to share the raw assets that they captured through the digitization process or project period. Those assets that fell into the "Other" category included papers, presentations, code, project websites, map layers and story maps, questionnaire and ethics protocols, and metadata records.





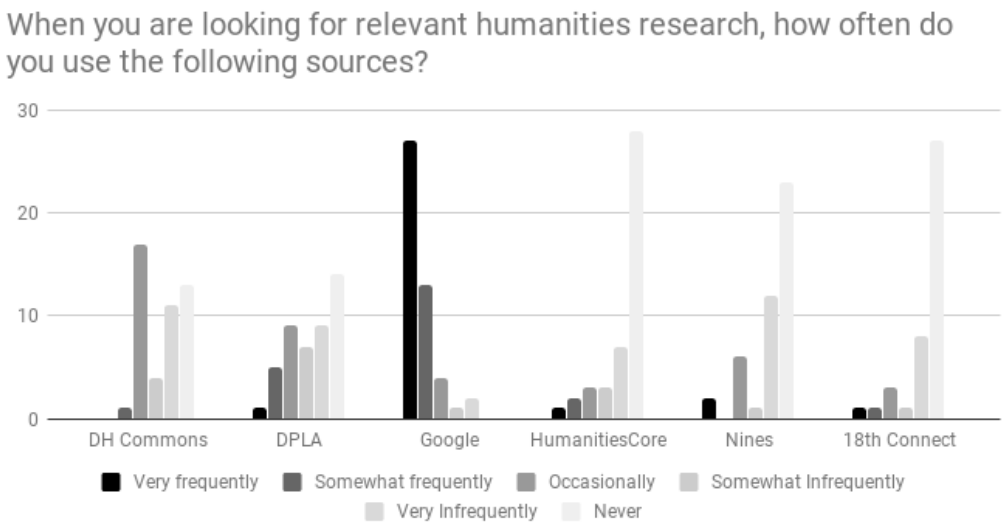
Results indicated that the majority of respondents shared their assets on GitHub and personal websites. They also used Institutional repositories and networked drives in significant numbers to share the assets of DH projects. The "Other" category resulted in a number of software and online tools, including ArcGIS online, Google Drive, project and institutional websites, and collection management software and tools. Respondents most commonly shared their final assets on their personal websites [n=10] and institutional networked drives [n=8]. They most frequently shared raw assets on GitHub [n=11] and personal websites [n=9]. Finally, they most commonly shared analysis scripts on GitHub [n=13], institutional networked drives [n=4], and personal websites [n=4].

In comparison to which assets an individual has publicly shared, the answers about where the assets are shared indicate that raw assets were the most frequently shared [n=50], followed by final assets [n=40], analysis scripts [n=28], and other types of assets [n=21].

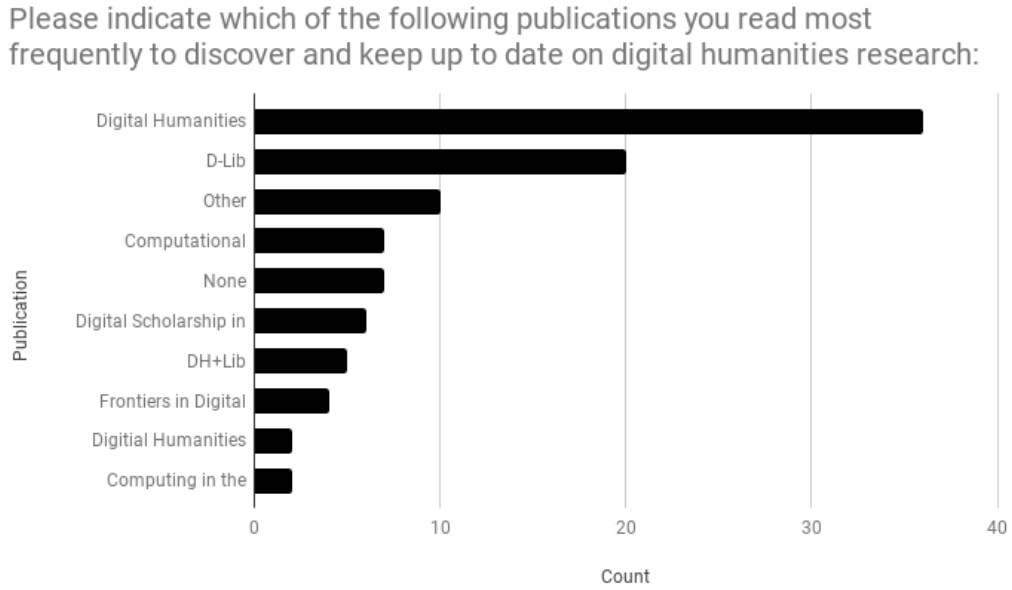


Searching and Discovery Habits

Investigating how scholars search for and find relevant DH research helps facilitate the appropriate application of taxonomies into system development and also identifies priority locations for metadata harvesting. To uncover this information, we asked a series of questions related to the frequency with which respondents use existing DH-related search engines, databases, journals, and registries, as well as questions related to their preferred ways of searching for DH scholarship.

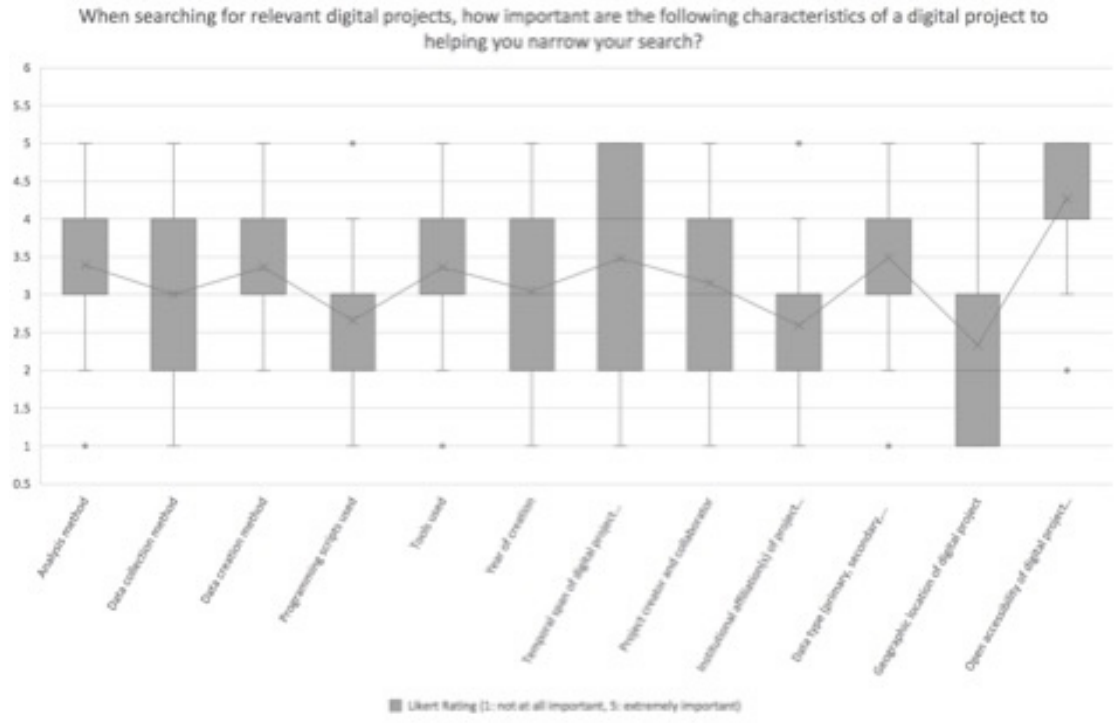


We provided the respondents with a list of six publications that publish DH scholarship, a "none" option, and an "other" option with the ability to add free text. Respondents most heavily used Digital Humanities Quarterly (36%), followed by D-Lib (20%), and Computational Linguistics (7%). The Other category resulted in respondents writing in 13 different publications. We included the following, which had a count greater than one, individually in the chart below: Digital Scholarship in the Humanities, DH+Lib, and Digital Humanities Now.



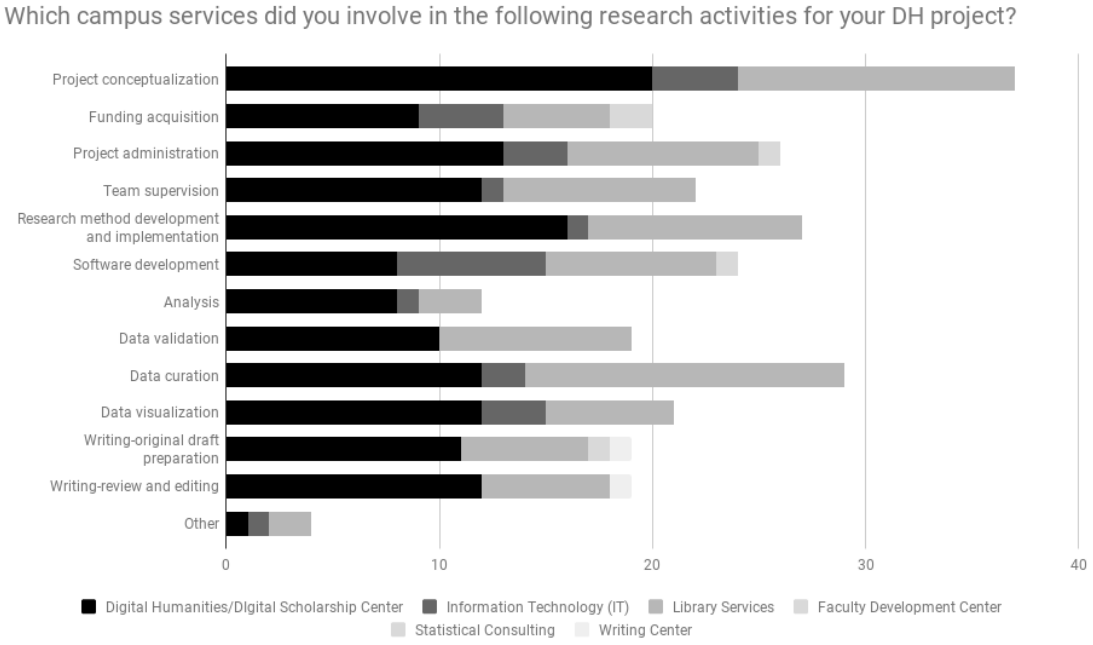
We asked respondents to rate twelve different methods by which they might search, discover, facet-over, or filter relevant DH scholarship on a five-point Likert scale ranging from “not at all important” to “extremely important”. The search methods that respondents rated as “extremely important” or “very important” were open accessibility of digital project (82.6%), temporal span of digital project (52.1%), data creation method (transcription, translation, recording, imaging, etc.) (48.9%), and tools used for project creation (46.6%). Methods that they rated most frequently as moderately or slightly important included institutional affiliation (63.6%), programming scripts used (59%), year of creation (58.7%), data collection method (53.3%), geographic location of digital project (53.3%), analysis method (52.3%), data type (primary, secondary, analyzed, raw) (51.1%), project creator and collaborators (48.9%), and tools used for project creation, collection, analysis, etc. (46.6%).

The figure below shows the distribution of responses coded as 5 for “extremely important,” 4 for “very important,” 3 for “moderately important,”, 2 for “slightly important,” and 1 for “not at all important.” The trend line follows the median responses for each search method.



Institutional Resources

We also asked respondents to indicate which six institutional services they used in their DH projects across thirteen stages of the project life cycle. They used digital humanities and digital scholarship centers most heavily for the following stages: project conceptualization (54%), funding acquisition (38.9%), project administration (44%), team supervision (50%), research method development and implementation (58%), software development (38%), analysis (53%), data visualization (53%), writing-original draft preparation (58%), and writing-review and editing (61%). They used the library most heavily for the data validation (42%) and data curation (46%) services related to their DH projects.



Conclusion

Digital humanities scholarship is multimodal and heterogeneous

The survey results overwhelmingly indicate that DH research is not bound by any particular method, analysis, or tool. Looking across the 42 DH projects we profiled, respondents used over 100 unique tools,

most of them for only one step in the research process within one distinct project. This complicates making any generalizations about DH workflows. Rather, the results of this survey support the claim that DH scholarship is incredibly varied in its approaches and development.

Reviewing the identified disciplines in which the survey respondents work, we can further see that DH scholarship is not limited to humanities fields but extends into the social science areas of sociology and anthropology.

Data availability is of significant importance for DH researchers

Although we provided twelve distinct methods as mechanisms for searching and discovering DH scholarship, the open accessibility of digital projects was of the most importance for survey respondents. Facilitating this type of discovery involves not just searching digital assets according to their accessibility but also ensuring the proper licensing for reuse.

Drawing upon our previous experiences with aggregating metadata at scale with SHARE, we recognize that each of these search methods, while seemingly easy, is incredibly complicated. While 67% of survey respondents indicated that metadata was an output of their DH scholarship, fewer than 35% of those used Dublin Core metadata elements. While Dublin Core aids in discoverability and is useful for general description, it does not include many of the values required for the searching preferences identified through this survey. Additionally, metadata values may be applied inconsistently, have missing information, or otherwise be of poor quality. Before many of these search methods are adopted, much work must go into cleaning, normalizing, and enhancing the metadata. While this remediation of the metadata often requires human intervention, advances in natural language processing, machine learning, and artificial intelligence offer promise as an automated mechanism to complete this curation.

Digital Humanities and Digital Scholarship Centers are integral partners throughout the DH workflow

The frequency with which survey respondents leveraged digital humanities and digital scholarship centers at their institutions for a wide variety of services was impressive. Respondents cited their use of these centers most frequently in all but two of the twelve DH project stewardship activities that we identified. In effect, these centers have become support mechanisms for faculty and researchers throughout the research life cycle. From support for conceptualizing DH research to consulting on analytical approaches through writing and finally publishing the outcomes, institutional digital humanities/digital scholarship centers offer a wide portfolio of services.

Where these centers live within an institution can vary widely, with some located in university libraries and others embedded in academic departments. Future research would benefit from better understanding if, how, and to what extent the organizational structure influences the services, support, and faculty interactions at these centers.

Limitations

While the results of this survey are enlightening, they are not without their limitations. First, given the online nature of the survey and the modes of distribution, it is very likely that selection bias had an impact on who responded. Many of the responses indicate a high use of technical tools to conduct DH scholarship, and it is unclear how representative this sample of responses is of actual practice.

Acknowledgements

We wish to thank John Russell, digital humanities librarian at Pennsylvania State University, and Elizabeth Waraska at ARL for reviewing, editing, and reading this report.

Appendix C: Workshop Agenda and List of Participants

Agenda

Code of Conduct

Location:

Four Seasons Hotel Denver
1111 14th Street
Denver, Colorado 80202
Cottonwood Ballroom A

DAY ONE

Noon Registration

1:00 Introduction and Welcome

1:30 Panel: Problems to Solve

- **Nikolaus Wasmoen**, Visiting Assistant Professor in Digital Humanities, University at Buffalo
- **Quinn Dombrowski**, Digital Humanities Coordinator, University of California, Berkeley
- **Annie Johnson**, Library Publishing and Scholarly Communications Specialist, Temple University

2:30 Break

3:00 Work Session #1: Use Cases and Value Propositions

4:15 Full Group Discussion & Debrief

7:00 Dinner

DAY TWO

8:00 Breakfast

9:30 Panel: Towards a DH Dashboard

- **Kathleen Fitzpatrick**, Director of Digital Humanities, Michigan State University
- **Jeffrey Spies**, Co-Founder & CTO, Center for Open Science

10:15 Break

10:30 Work Session #2: Sketching Solutions

11:30 Full Group Discussion & Debrief

12:00 Lunch

1:00 Work Session #3: Getting it Done: Workflows and Partners

3:30 Final Group Discussion

4:00 Workshop concludes

In the original project proposal, the workshop was geared toward pedagogy—specifically the opportunity to bring Postdoctoral Fellows from the Council on Library and Information Resources (CLIR) program[1] and SHARE Curation Associates[2] together as overlapping networks of expertise and learning. When the Curation Associates program did not receive funding beyond its pilot phase, the team decided to meet this basic objective by including a critical mass of CLIR postdocs and former Curation Associates in the event—along with CLIR program staff—as we began the requirements-gathering process. Pedagogy was an important discussion theme throughout the workshop, in terms of identifying skills library staff need to support DH scholarship, the importance of supporting DH teaching, and the use of public goods like SHARE in both of those instances.

List of participants

Laurie Allen

Scout Calvert

Corey Davis

Rachel Deblinger

Susan Doerr

Quinn Dombrowski

Ixchel Faniel

Matthew Harp

Cynthia Hudson-Vitale

Annie Johnson

Rick Johnson

Joan Lippincott

Nancy McGovern

Paige Morgan

Jeremy Morse

Thomas Padilla

Joanne Paterson

Megan Potterbusch

Wendy Robertson

Barbara Rockenbach

John Russell

Judy Ruttenberg

Emily Sherwood

Jeffrey Spies

Lisa Spiro

Elizabeth Waraksa

Christa Williford

Micah Vandegrift

Endnotes

1. “Postdoctoral Fellowship Program,” Council on Library and Information Resources, accessed November 14, 2019, https://www.clir.org/fellowships/postdoc/.
2. “Curation Associates,” SHARE, accessed November 14, 2019, https://www.share-research.org/tag/curation-associates/.

Appendix D: Focus Group Report

Introduction

In May–June 2018, the project team visited six US and Canadian universities and held focus groups comprised of DH librarians, scholars, and DH center staff. The project team selected the six campuses in consultation with the project advisory board and with the 30 participants of project workshop, held that February. Criteria included (1) presence and location of the DH center (in or outside the library), (2) geographic diversity, (3) presence of a CLIR Postdoctoral Fellow, (4) reputation for innovation in DH, and (5) inclusion of institutions that are less frequently profiled. In addition to facilitating the focus groups, project leaders emailed contacts at the six institutions, supplied email text to invite participants, a registration form, and offered a small budget for catering.

Methods

Once the sites were confirmed, the project team signed up to facilitate them in pairs, and finalized the questions, which were designed to mirror the structure of the project survey: with questions about assessment, challenges & resources, stewardship, and solutions. (Focus group script attached). The questions also drew from the survey results, and built on discussions from the workshop.

IRB exemption was granted through Washington University in St. Louis, where Cynthia Hudson-Vitale was employed at the outset of the award. The exemption was granted on the basis that the questions asked for facts, not opinions, and that the research was classified as "non-human."

The focus groups were recorded using bi-directional microphones with USB connections to laptops. The recordings were transcribed by 3PlayMedia and the digital recordings were then destroyed. The project team analyzed the questions manually, extracting themes and summarizing comments.

Penn State University Libraries intern supervisor Heather Froelich provided an additional analysis of the text, focusing on high-level themes.

Challenges & Resources

Tools and technologies for FAIR (Findable, Accessible, Interoperable, and Reusable) digital humanities

Key themes: *adopting and adapting existing best practices; discoverability through professional networks and university networks; interventions from the department side make things happen more often than not; RDF (Resource Description Framework), linked data, etc. don't scale well*

Focus group participants by and large indicated the importance of metadata. Many create metadata of various kinds for their projects and, while there is a realization that it is needed for discovery, metadata creation is considered onerous, time-consuming, and too often left as an afterthought. Projects are often inventive and unique, so which description, which classification standard should be used? This is not always clear-cut. A wide variety of standards are used, from Simple Dublin Core and TEI (Text Encoding Initiative) to more specialized types like ArchaeoCore[1] and VRA Core.[2] Further, because ontologies can be biased, because words matter ("this is the humanities after all!"), and labels can be burdened with cultural bias, some find it hard to agree on which ontology, which standard to apply, and may create something entirely new. A lot of time is spent creating boutique ontologies, which means then that they are not interoperable, defeating the goal of standardization. There may be a role to play here for those librarians and archivists who specialize in the application of metadata with setting up best practices and training and teaching others. It may be that there is a standard set of metadata that aids in sharing, used in conjunction with a second, more in-depth ontology that suits the needs of the material better. One focus group participant mentioned that they have begun teaching students in their research methods course about metadata and its crucial role in discovery and noted that once students

realized its power, they were more inclined to pay attention to it. In one particular case, funding for DH projects is approved by the dean, who refers the team to the library where plans for metadata are discussed. DH projects could benefit from data management planning and the library can play a role here.

Challenges in making it FAIR

Key themes: *tenure and promotion; evaluation, assessment, money (who, how, and when)*

While this question was meant to express the full extent of challenges to making DH projects FAIR, focus group participants predominantly focused on the challenges of evaluating DH projects and the financial aspects of sustaining a project beyond the grant period.

One of the biggest issues identified by participants was related to promotion and tenure, specifically, how DH contributions and projects have not been systematically integrated into departmental policies and guidelines for tenure. While some professional organizations such as the Modern Language Association[3] and the American Historical Association[4] have developed evaluative guidelines for digital scholarship and DH, these guidelines have not been widely implemented at the institutional or departmental level and only cover a small portion of the disciplines that currently develop DH projects. This produces little incentive for faculty who are seeking tenure to be involved in DH projects outside of the principal investigator role and more generally outside of the duration of the grant award period.

Another challenge raised in discussions was the lack of recognition or reward for many parts of the DH production process and ongoing maintenance (such as code, metadata, digital surrogate creation, etc.). Given that there are a variety of roles and contributions a researcher or staff member may make towards a DH project, few roles, other than principal investigator, are visible or widely acknowledged. This reduces incentives for participation as their involvement would be largely unrecognized.

The financial aspect of developing and maintaining DH projects was also a common thread among focus group participants. While researchers can receive funds to develop a DH project from a funding agency, monies are often lacking for ongoing maintenance and preservation of assets. We heard from various focus group participants that oftentimes the library is approached at the time that the grant is sunsetting and faculty are looking for ways to keep the project or website functioning. This results in a complex evaluation by the library, comparing the existing technology stack and expectations for contributions with library resources and skills. We heard repeatedly from focus group participants that the earlier a researcher could think about sustainability beyond the grant period—especially if they were considering working with an institutional digital scholarship center or the library—the more likely it will result in a positive outcome.

Dissemination

Key themes: *social media; traditional publishing; altmetric*

The dissemination of DH work depends somewhat on the output of the project. While some rely on building in search-engine optimization for a website, others use Twitter to share news about a database or code on GitHub. Still others link to their ORCID profiles as a means of sharing projects. It appears that old-fashioned networking is a popular means of dissemination, as DH-ers will email and tweet to colleagues about their work, and look for email and association lists to send notification of teaching materials and projects to those they think are interested. How and why to disseminate depends upon the value placed on the piece of scholarship. Much of the resultant work is experimental, ephemeral, and becomes orphaned. Once a project has reached its conclusion, or researchers move onto other projects, focus moves to the next new thing. As reflected above, not all parts of the DH project are valued equally, especially in terms of career advancement. Overall, the article still remains the primary vehicle for dissemination and evaluation as it can be included in promotion and tenure dossiers.

Stewardship

The stewardship of digital humanities projects is not limited to any one group or organization across an institution. An analysis of the focus group transcriptions across six institutions indicates that decisions regarding who stewards a project are primarily driven by who has the skills or expertise, what the scope of the project is, and the overall budget. A number of the focus group participants across institutions mentioned the role of the library in stewarding and supporting digital humanities projects, while grants coordinators and support were also widely cited. Critical examination of the transcriptions also suggests that much stewardship is done on an ad hoc basis, some faculty respondents indicated the difficulty they had in finding the resources and people to help them complete their scholarship. One method to overcome this involved forming a small team of individuals with representation from the library, subject liaisons, DH centers, campus IT, and high-performance computing to review DH-related project proposals as they came in. This small team would then determine what support they each could provide.

Resources on campus

Key themes: *no coherent group of people; levels of support; who knows about them; limited by skills, expertise, scope, and budget; who can learn something quickly; done at a human scale; never one-size-fits-all approach*

The ad hoc nature by which many institutions handle digital humanities projects is reflected in the highly specialized workflows that were described for the stewardship of projects. Very often DH project teams are formed based on the nature of the project and the skills required to manage the technology and needs of the project. It was noted that while web archiving is taking hold at many institutions, dealing with DH projects that live on WordPress, Drupal, or any database were challenging to steward. One focus group participant noted, "Most of the projects are not just a bunch of digital objects as output, they're these complex technology stacks, and we don't really have good solutions for that in libraries."

This also relates to another challenge on how to make decisions about digital humanities projects that use outdated or obsolete technology. Focus group participants reflected on the complexities in determining whether or not to remove a website, to migrate it, or simply let it persist. Given the speed at which technology refreshes and the potential for sites or embedded objects to break, one participant mentioned making PDFs or screen shots of digital humanities projects as a method to capture content in context.

Funder or institutional mandates for sharing

Key themes: *haphazard; make the most of what is available; if it is done, it falls to the departments that have the time, scope, and money (both within and outside the library)*

With respect to meeting funder or institutional mandates for sharing data or research outputs, we heard from focus group participants that the requirements are often met in a haphazard manner and using existing tools and infrastructure. When determining what institutional support can be provided very often it comes down to capacity of time, scope, and money. Not surprisingly, if funds are available to support the project, there is often more capacity to take on a project.

Many of the focus group participants felt it would fall upon funding agencies such as the NEH to dictate best practices for scaling DH projects, including guidance on large-scale production and stewardship throughout the research life cycle. Respondents felt that the NEH could also easily support the development of standardized workflows for certain types of DH projects that are not highly complex, such as Omeka sites.

Gaps in DH stewardship services within the library

Key themes: *no real standards; very cobbled together; make it work at your institution*

A theme around dependencies also surfaced throughout the focus groups, specifically around how to ensure the ongoing availability of

DH projects. Many participants noted the complexities of stewarding maps given the proprietary nature of the most prevalent mapping tool, ArcGIS. While web archiving tools can scrape much information from webpages, maps and databases are a known limitation. Additionally, there were multiple discussions about stewarding any code or software that is produced as part of a digital humanities project. While tools like GitHub can store and make the code discoverable, tools and workflows for ensuring any dependencies are available or just developing through emulation services and containerization strategies.

Assessment

Key themes: *tenure and promotion; what counts; hard unless it is in a peer reviewed paper; where does a project go?; what is the output? impact factor → how is this baked into outreach; no vision for what impact factor would look like (they have lots of analytics, but don't know what to do with them)*

Assessment appears to be an area of potential growth for many of the organizations. While a recognition of the value of assessment is established, in practice each of the focus groups demonstrated a lack of clarity, resources, and standardization for assessment. Web statistics (downloads, hits, pings, clicks, and referrals) are still in use but there is also a desire to show impact on reuse and integration into pedagogical application and integration within the scholarly record. (In other words, is a given project being cited and used in other projects and publications?) Another point is that projects are being measured by their level of completeness and whether they are small "rapid" projects or ongoing long-term. Assessment may evaluate projects throughout their processes.

Of specific value are indicators of reuse and learning. Measures include peer reviews in the form of advisory group feedback, personal stories from students, faculty classroom use feedback, extracting URLs from citation references, and the integration of projects into library management systems (LMS) and tracking their pings. One participant felt strongly that measuring projects before their integration and

indexing within discovery systems is premature because the full extent of use is not possible until the projects become discoverable. They also noted that external changes can drive changes in metrics because things must be available, discoverable, and accessible before they will be used, and outcomes will change over time.

Metrics

Key themes: *we have some good models, but few and far between*

Outside of web statistics, metrics and methodologies for assessment are nascent. Having proposals accepted by conferences or published in journals based on project work were indicators of success. Some professional societies, such as the Modern Language Association (MLA), have issued guidance on the evaluation and assessment of digital scholarship.

The SHARE project may be able to consider investigating partnerships with groups like the American Historical Association (AHA) and the MLA whereby SHARE could improve discovery of semi-official, blog-style resources that provide project feedback and help evolve the use of DH projects in education courses. Standardized metadata tagging and criteria could facilitate assessing project completeness and provide categorization opportunities. Adoption of pedagogical approaches beyond a DH project team may be assessed if projects and methods are cited in other projects and publications.

Some of the conversations seemed to drift from the root topic of metrics and assessment, such as at the University of Victoria, where the subject was largely bypassed, or at Columbia University, where time was spent discussing the difficulties of measuring web archives. While assessing web archives may be out of scope for this project, estimating the long-term value and use of older projects is indeed within the spirit of persistence of information access, reproducibility, and discovery.

Solutions

Reuse

Key themes: *good enough to borrow; open source software; digitized resources*

While the scholars represented in the focus groups emphasized the high-touch, personal networking nature of their DH communities (Twitter lists, email lists, conference networking), collectively, they could envision a discovery dashboard for finding DH work. Participants identified desirable facets for such a dashboard, including: image, text, tool, method, digital collections, thematic research collections, GIS component, name, funder, and language.

There was a lot of discussion of trustworthiness and certification of text corpora, as well as the need for credit for the scholars who created the aggregations, and provenance of their sources, if these are to be the basis for text mining and analysis. Similarly, participants expressed desire for rating the trustworthiness or replicability of the analysis. This too seemed to be a heavily manual part of the work—contacting people directly to determine everything from provenance to reuse rights. There was some speculation that since "everyone" knows about Creative Commons (CC) licenses, the absence of licenses indicates apathy toward reuse. Specific tools that people expressed optimism for were Omeka (a form indicating use), Wikidata, and TAPoR. Several groups surfaced the very basic challenge of creating bibliographies in DH work, from a dearth of tools and standards to lack of interoperability among them.

Ideas for discovery

Key themes: *metadata; DiRT (Digital Research Tools); DPLA (Digital Public Library of America), HathiTrust; build a space that combines Humanities Commons and GitHub; much disciplinary siloing*

With respect to finding, using, and reusing software tools, there were interesting comments about the need for easy to use tools (such as

intuitive interfaces) for DH analysis on the one hand, and the need for scholars to understand the tools they use (algorithms, etc.) and for peer reviewers to understand those same issues in order to properly evaluate the work, on the other hand. This led to general comments about interdisciplinarity and the need for mathematical and statistical expertise in some DH projects.

When discussing scaling and the discovery of DH projects many focus group respondents indicated that these projects often live in a vacuum at their institution and it is very difficult to think about scale outside of their own institution. Even at the local level, DH production teams struggle with discoverability.

Conclusion

Participants at each institution provided unique use cases that exemplified the same challenges digital humanities projects encounter. While the website has become a staple of the digital project, it is still not considered as highly as the scholarly article, which may merely summarize the value and impact the web projects provide.

All participants agreed, while not able to necessarily articulate solutions, that better metrics and assessment such as citation counts, or notifications of reuse as pedagogical instruments, that go beyond visits and download counts would help evolve websites beyond their marginal affiliation. Getting metrics involves deeper integration with library and archival discovery systems, mechanisms to standardize metadata schemas, and notification systems to enable peer review, much in the same way preprint publication preregistration systems and preregistration facilitate peer review for scholarly literature (articles) and scientific research methods. Like a snowball gathering speed and weight, greater discovery leads to greater use, fostering increased author citation/credit and reuse of work that expands and mobilizes digital humanities knowledge.

These reflections speak to the outcomes for which SHARE was designed to facilitate: aggregation and categorization of products, people, and organizations.

Focus Group Survey Instrument

Introduction

Thank you for participating in this interview. This study is conducted by the SHARE team in collaboration with the Association of Research Libraries and funded by the National Endowment for the Humanities. SHARE is a database of open metadata aggregated from over 100 sources, including Humanities Commons, PhilPapers, and Papyrus. We are investigating the requirements to link distributed digital humanities scholarship so that the scholarship is more findable, accessible, interoperable, and reusable (FAIR). The purpose of this interview is to better understand the stewardship of digital humanities scholarship. You have been asked to participate because we have identified that you are a creator or steward of digital humanities scholarship.

We estimate that the interview today will take approximately one and a half hours. Aggregated, coded answers will be shared but your name will not be associated with any of your comments. Any directly identifying information will be redacted. Audio recordings will be destroyed after coding. If you prefer not to answer any questions, let us know and we will skip them. You may end this interview at any time.

Do you have any questions before we begin?

Challenges & Resources

1. A workshop identified two main obstacles to creating structured, harvestable, and citable metadata: (1) DH projects tend to be independent websites, rather than publications; and (2) Project workflow can privilege the website presentation and design over underlying data model.
    a. What tools, technology, and resources have you used to make the outputs of your scholarship more findable, accessible, interoperable, and reusable (FAIR)? [For example: metadata creation, human resources, technical infrastructure, university services and support]
    b. What challenges have you encountered in making the DH project FAIR?

Stewardship

1. What university support have you leveraged to complete your DH project? (statistical consulting, writing center, library services, etc.)
2. Research libraries may have workflows for handling web archiving, software preservation, research data management, and publications. What DH products fall outside of these buckets?
3. How are you currently addressing funder or institutional mandates for research data management and data sharing?

Assessment

1. Thinking about a specific DH project you have created, how was the project assessed (in terms of quality, impact, reuse, etc.)?
2. What, if any, metrics did you collect to assess the project?
    a. How did you collect those metrics?
3. What non-metric based assessments did you use to assess the project? (personal stories, contextual information, etc.)

Solutions

1. What components of other scholars' work have you reused? Examples: published work, tools, software, digitized collections
    a. How have you reused these assets?
    b. How have you indicated to the project owner you reused the assets?
2. What user interface, discovery, or facets are required to make a discovery dashboard appropriate for DH?
    a. Is there a unified language, taxonomy, or metadata standard that describes the methods, components, and distributed location of DH assets currently?
        i. If yes, how frequently is it used or applied?

Conclusion

Is there anything else you wish to share with us about the discovery and stewardship of DH research and workflows? Have I not asked any important questions?

Endnotes

1. ArchaeoCore, accessed November 14, 2019, http://www.ifaresearch.org/archaeocore/.
2. VRA Core, accessed November 14, 2019, https://www.loc.gov/standards/vracore/.
3. "Guidelines for Evaluating Work in Digital Humanities and Digital Media," Modern Language Association, last revised January 2012, https://www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Evaluating-Work-in-Digital-Humanities-and-Digital-Media.
4. "Guidelines on the Professional Evaluation of Digital Scholarship by Historians (2015)," American Historical Association, accessed November 14, 2019, https://www.historians.org/jobs-and-professional-development/statements-standards-and-guidelines-of-the-discipline/guidelines-on-the-professional-evaluation-of-digital-scholarship-by-historians-(2015).

Appendix E: SHARE-ing Omeka in the Web of Digital Scholarship

1. Introduction[1]

The SHARE project[2] needed a method to extract descriptive information from a wide range of Digital Humanities projects. The DH community has widely adopted Omeka CMS to create DH projects. Users of Omeka can organize a wide range of digitized artifacts as items into one location and then showcase them on a public website, typically within a collection or exhibit. Omeka uses items with metadata as the main building blocks to create projects. Omeka's website has a lengthy list of published projects, but it only lists each project's title along with a one-sentence description.[3] The only piece of extra information is a short list of plugins used to create each site. This dearth of information affects the discoverability of these projects by those outside their institutions and this list. Omeka's only public export option captures either items or collections as JavaScript Object Notation (JSON) embedded into a browser window. All other export options require administrative access to a project. This forced the SHARE team to find tools that could extract information from published websites. We tested two web scrapers to accomplish this: ParseHub,[4] a downloadable application, and Webscraper.io,[5] a Chrome browser extension. Both have free options as well as plenty of documentation and forums to help with any issues that arise.

2. What Can Omeka Do?

As a first step, we gathered a wide range of Omeka projects to load into web scrapers in order to see what kinds of projects exist that utilize the system. This also allowed us to test different organizational styles between projects. The sites we found for testing generally fell into four categories: exhibits based, collections based, map based, and sites for coursework. Exhibits-based projects, such as the Florida Memory project[6] and Making Modern America,[7] hold multiple separate projects on a common theme. They are the most content heavy and the most dense in terms of the web scraper sitemaps necessary to extract the information contained within them (see Appendix A §2 Creating a Sitemap). The collections-based projects tend to have the most

individual items grouped into one category along with a description of that grouping. One example is Farms to Freeways,[8] an oral history project based in Australia. Map-based projects are the most difficult to scrape because of their visual nature. Maps typically use pop-up windows to show items with pertinent information about particular map locations with little to no content. Map-based projects usually relegate context for the map to an about page, which usually holds information regarding the overall project. More often than not, there is no collections or exhibits page. One of the more well known of these is Histories of the National Mall.[9] The last major group is sites for coursework, which tend to be used to teach how to create a DH project rather than to present information to the wider DH community. The 9/11 Living Memorial Digital Archive is one example of a site for coursework.

For the purposes of this project, we decided to focus on collections- and exhibits-based sites because they tend to be well-developed and have enough variety to test how the different elements within them affect scraping. When we were searching for test sites, they were by far the most common, as roughly 9 out of 10 projects fell into these categories. The two projects we used most often for testing were Colored Conventions[10] and Florida Memory,[11] both of which utilize a wide range of Omeka's capabilities. Colored Conventions highlights the organized meeting of free Blacks during the 1800s and is centered around the meeting notes and documents from these conventions. Most of these are organized into individual items, typically one for each convention. It also has a number of exhibits discussing topics from the conventions. Colored Conventions tested how the web scrapers tackled a site with straightforward groupings of collections and exhibits. We needed a more complicated site to see how the scrapers dealt with complexities. The Florida Memory project is the main website for the State Library and Archives of Florida. It organizes its information into specific types of physical media, like photographs, videos, and audio. Each of these have their own exhibits and collections. Other projects tested, like Gothic Past,[12] use a mix of both collections and exhibits and tend to have fairly simple HTML.

3. HTML Scraping Tools

After collecting a group of websites, the project switched focus to the tools used to extract information from them. One of the most basic ways to do this is to pull information directly from the web pages published by the projects. Several tools called web scrapers exist that perform this task. We identified two to test the scraping of Omeka projects. Both satisfied the exportable, flexible, and scalable requirements of this project. The scrapers also needed to have low barriers of entry and be fairly easy to use so that those with limited resources or technical expertise could still utilize them for their projects. They also needed to scrape a wide range of different projects that have unique organizational structures. Another positive feature, which also improves scalability, is the ability to scrape multiple languages. It is also important to be aware of any copyright information listed on a website prior to scraping it. Consult a copyright lawyer or professional to ensure the information can be legally scraped and/or shared.

The two web scrapers tested here use HTML to identify different components of each page. They use sitemaps as umbrellas to hold all the information about each project, from the homepage URL to the final scraped data for export. Sitemaps serve to organize all information about scraping a website in one location, housing all the project information as well as selectors. The selectors perform a wide variety of actions, from clicking links on a page, to scrolling, to telling the scraper to select and extract individual pieces of information.

3.1. ParseHub

We tested ParseHub first because of forum comments that stated it was an easy-to-learn and use system. ParseHub is a free computer application that can be used on either a Mac or a PC. It requires setting up an account with an email and password prior to use, after which sitemaps can sync with the ParseHub servers for backup and sharing of information. This can be helpful for group projects, which can be publicly viewed if using the free version. The system is very easy to

use and is a good starting point for those who want to learn how to set up sitemaps for web scraping. It uses the HTML code on a web page to select elements from which to scrape information. ParseHub uses sitemaps as the main method to scrape a website. Figure 1 shows the large panels users view to easily find different components of the web scraper.

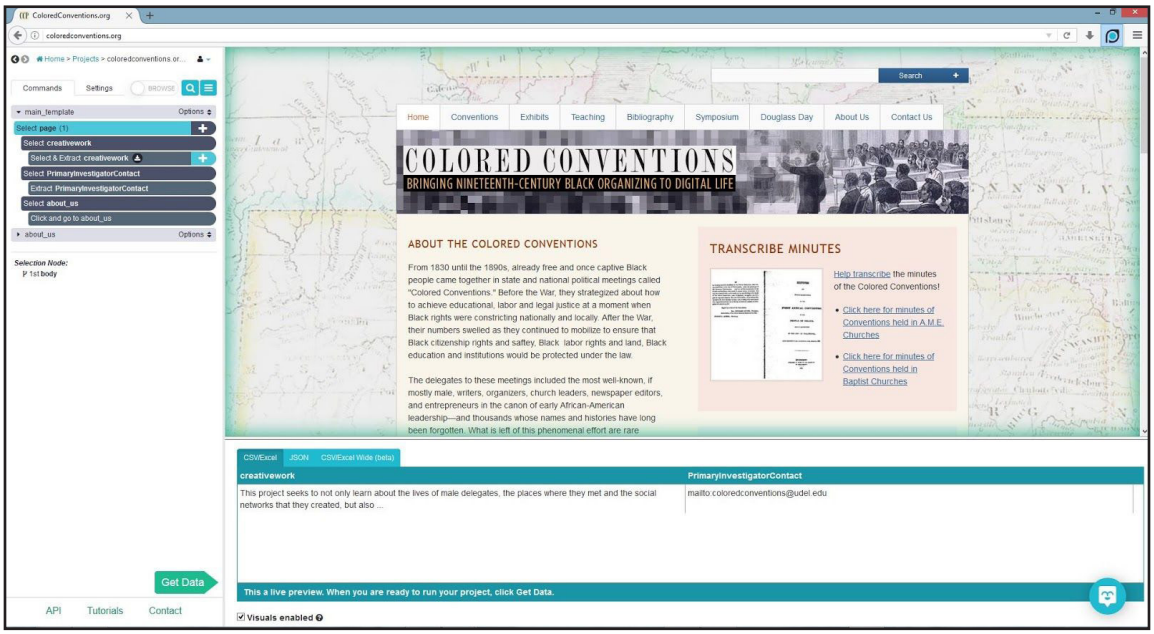


Figure 1—View of ParseHub when opening a built project

Everything in ParseHub is point and click, and the layout makes working in the system simple. The largest panel in Figure 1 is the web page that the selectors will interact with during a scrape. The selector panel on the left includes the sitemap, templates, and selectors for the entire project. This panel also has a 'Get Data' button at the bottom, which initiates test runs and scrapes (See §3.3.1. ParseHub Export). If a selector is clicked in the left-hand panel, the data held within it appears in the bottom panel, where users can switch their view of the data between CSV and JSON formats.

The + button in the selector panel shows the different types of selectors, such as select, click, scroll, and so on. Once you choose the selector type, the system highlights the appropriate HTML items in the main project window whenever the cursor hovers over them,

and users can click on the items in the project that they want. As Figure 2 shows, the Colored Conventions homepage has a selector for the title of the project (creativework), the contact information (PrimaryInvestigatorContact), and the about information (about_us).

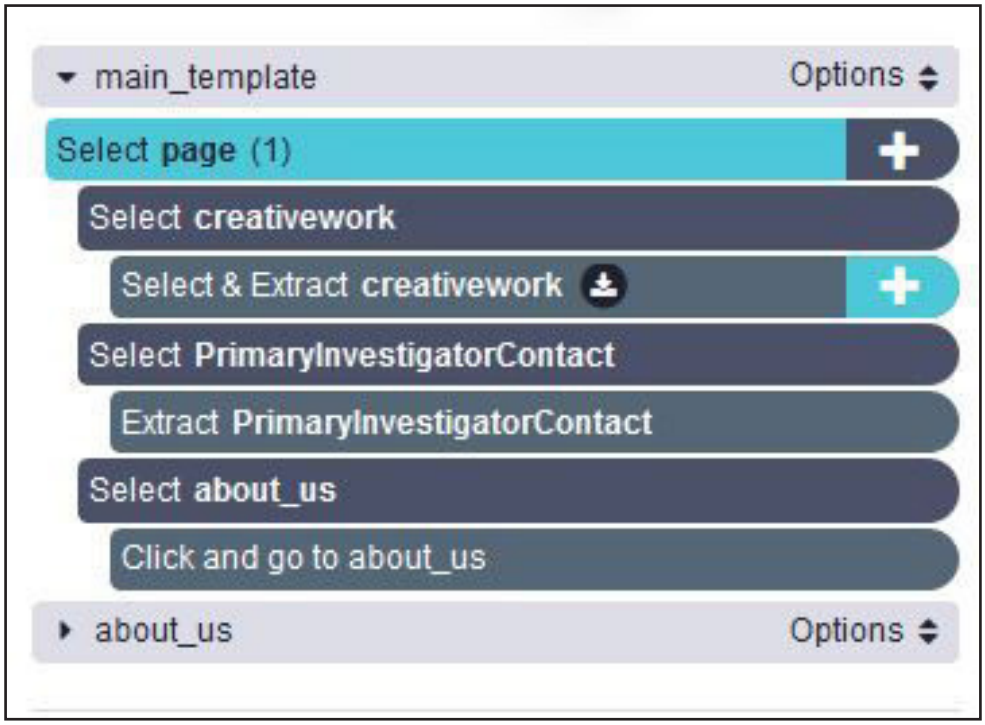


Figure 2—Sitemap for Colored Conventions in ParseHub

Separate selectors are needed for each link from which the user wants ParseHub to extract information. These selectors serve a second purpose within ParseHub: they tell the system how to navigate through the website during a scrape, such as moving from the homepage to an exhibits page, or identifying whether there are links to an organization that helps create the website. A link selector is also the last selector in the template, so the scrape can move on to another page. Because of this, each page has its own template in the sitemap. ParseHub gives a preview of the data for the current selector at the bottom of the app.

If there is more than one entry of the selected tag, such as a list of links, ParseHub highlights the entire list, as seen in Figure 3. ParseHub highlights the links that are in the surrounding tags in the HTML code in yellow.

Figure 3—Selecting lists

Any items that are green are set for extraction. Users can add the yellow highlighted links to the selector for scraping by either clicking on the link itself or the yellow check mark, both of which will add them to the sitemap for extraction (the X will not). In this way, users include only those items that are relevant to the project at hand. Once all are added, users can go back to the selector panel on the left side of the software to add them to the sitemap.

Once the sitemap is built, another pane in the left-side panel can either run a test or a real run, as seen in Figure 4. Test runs do not extract any data, as they are essentially trial runs to ensure that everything is set correctly within a sitemap. A test run cycles through the templates to ensure that each selector has an action attached to it, like extract or click, and to make sure it goes from one template to the next. It gives an error for those templates that are missing elements. The test run continues to run on repeat until stopped.

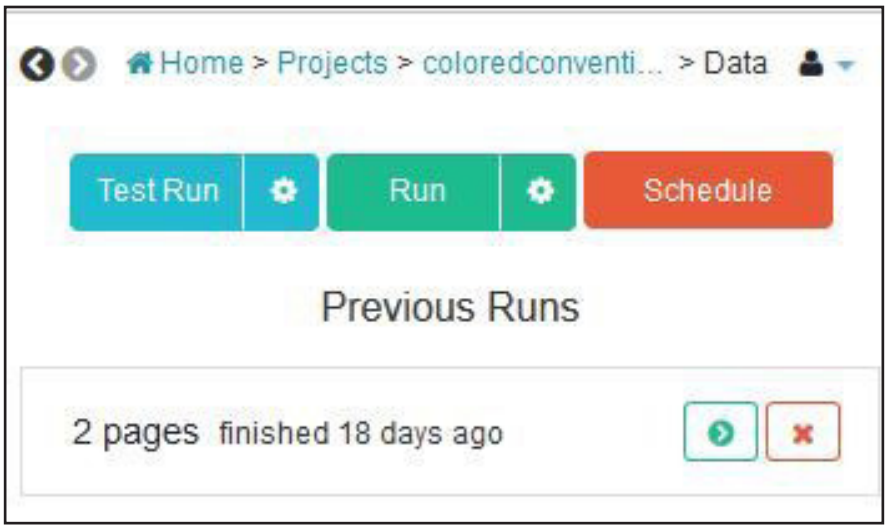


Figure 4—ParseHub allows for a Test Run prior to the extraction of information

In order to scrape information, users click the green 'Run' button, as seen in Figure 4, which then scrapes the website. Once completed, ParseHub sends an email to the address associated with the account stating that it is finished. The final data can then be exported as either a CSV or JSON file.

The biggest drawback of ParseHub is that the free version limits users to a total of five projects at one time. Once five projects are active, you must delete one before adding another. This is not an issue for someone who wants to learn how to scrape a website, but it is a severe setback for anyone trying to scrape multiple websites for a larger project. ParseHub also considers projects created with the free version to be public, so others can potentially viewed or search for them on ParseHub's servers.

3.2. Webscraper.io

The other scraping tool we tested was Webscraper.io. It is a free extension for the Google Chrome web browser. Unlike ParseHub, this web scraper utilizes the web browser itself as the basis for scraping (See Appendix A for a detailed Webscraper.io how-to guide). Users of Webscraper.io can extract information from any public website using HTML and CSS and then export the data as a CSV file. Unlike ParseHub, which loads the website into its own dedicated system,

Webscraper.io takes advantage of the developer tools available in Chrome to select the different elements directly on a website, including links, tables, and the HTML code itself. The scraper uses the developer tools to look at the code that generates everything on a web page, from text, to images, to the layout. Webscraper.io uses this code to select the different elements of a website to either extract information or to navigate through the website. There is a snapshot of a selector menu in Figure 5.

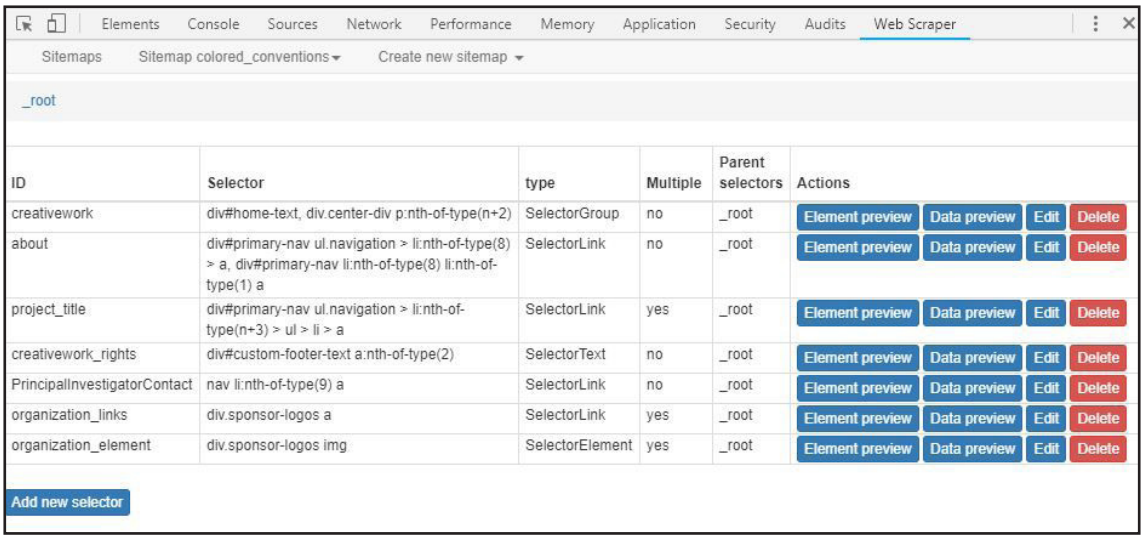


Figure 5—Webscraper.io selector list

The first column in the selector list is the name of the selector, which doubles as the column header in the exported data file (See §3.3.2 Webscraper.io Export). The second is the HTML and/or CSS code that the scraper will use to extract information from the web page. The third is the type of selector it is, whether it is a link, text, etc. The 'Multiple' column denotes whether the selector will extract more than one of the elements within the code, such as content paragraphs or links. The 'Parent selectors' column shows where the selector falls in the sitemap hierarchy (See Appendix A §4 Creating a Selector). The 'Actions' column contains buttons that users can utilize to preview information or to edit or delete the selector (See Appendix A §4 Creating a Selector).

Users can select multiples of the same type within the same selector. This is helpful for scraping lists of links or capturing several content paragraphs at one time. On a related note, Webscraper.io can also select more than one type of HTML tag to group them together. For instance, it can group a title and description that may each have different HTML tags together in one selector. There is also a selector called 'Group' that can group things together in the same cell of a CSV file, such as the paragraphs of a content page. This group selector includes the character \n, a marker that denotes the separation of tags with the website HTML, to separate the tags in the scraped data.

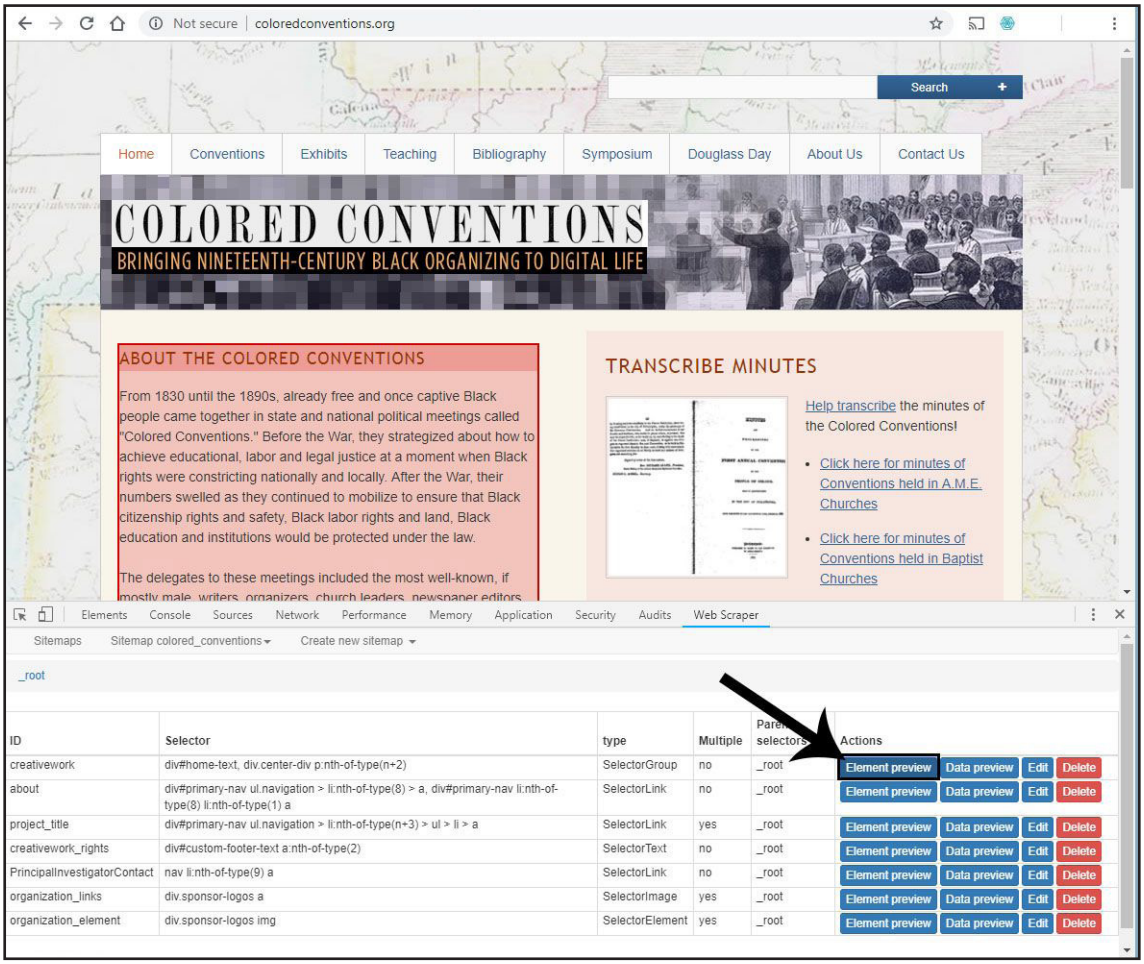


Figure 6—The 'Element preview' button creates a red highlight around all HTML elements entered into the selector

One useful aspect of all the selectors is that you can preview both the elements selected and the data they will extract. As seen in Figure 6, the element preview highlights everything on a page that the selector has in its code, ensuring that users have the elements they want. The data preview does the same thing for the information set for extraction. The exported CSV file for Webscraper.io includes both the name and URL for all links set in the sitemap, whereas ParseHub only includes the name.

Another helpful thing that Webscraper.io has but ParseHub does not is a selector graph, which is extremely helpful in understanding the hierarchy of selectors within a sitemap (See Appendix A §3 Selector Graph). The graph shows where a particular selector connects with others, as well as previewing how data will be organized in an export file (See §3.3.2. Webscraper.io Export). See Figure 7 for the selector graph built for Gothic Past.

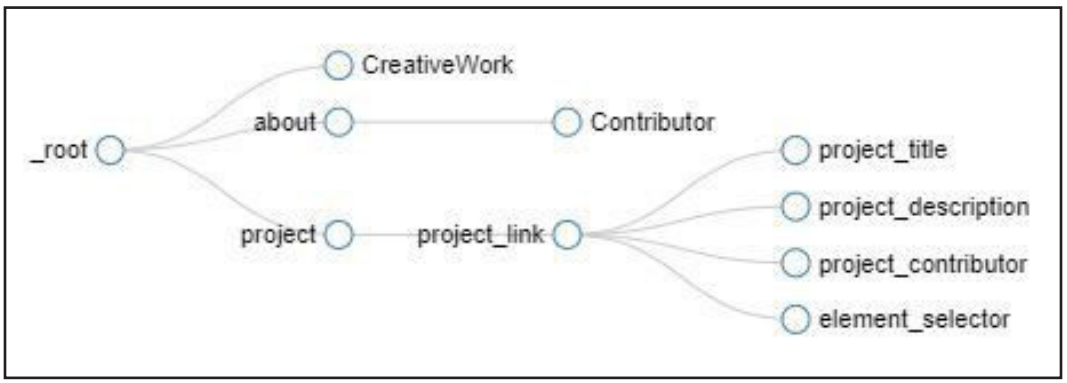


Figure 7—Selector graph for Gothic Past

You can see that the root page, what Webscraper.io labels the starting URL or homepage, has three selectors: CreativeWork (website title), about, and project. More selectors appear next to the about and project selectors with additional information for the scraper to extract. As a general rule, only selectors that extract links have more selectors listed beyond them in the tree. The graph shows how the data is organized within the scraper by highlighting which selector is a parent or child of another (See Appendix A §3 Selector Graph).

3.3. Exporting Scraped Data

Both web scrapers export CSV files, which can be loaded into virtually any system due to their data organization. Each selector name in each system becomes a column header when a user loads the CSV file into spreadsheet software such as Excel or Google Sheets. As detailed below, ParseHub and Webscraper.io organize their export files differently. ParseHub condenses information into blocks that correspond to the templates created for each page, while Webscraper.io keeps the hierarchical structure seen in the selector graph.

3.3.1. ParseHub Export

ParseHub CSV files have blocks of information corresponding to the different templates in the sitemap. The template for the homepage appears in column A and row 2 in the Florida Memory export example in Figure 8.

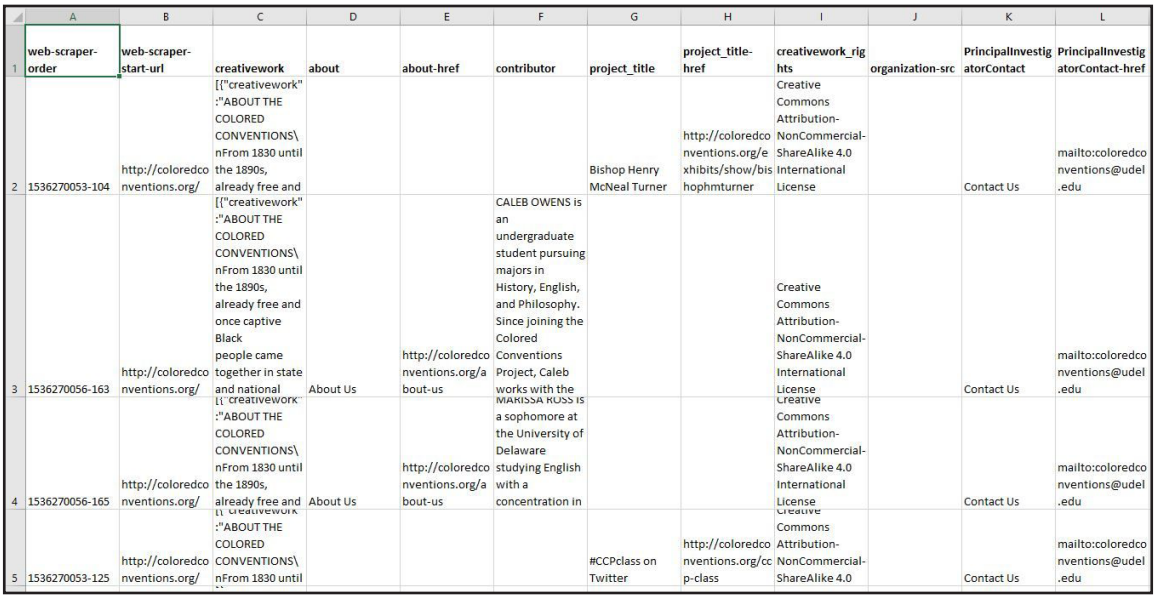
	A	B	C	D	E	F	G
1	about	description	collection_name	collection_url	collection_description	exhibit_title_name	on
2	About Florida Memory	The State Archives of	Florida Maps	https://www.floridam	Historic Maps of the		
3	About Florida Memory	-"-	Florida Folklife	https://www.floridam	Documenting Florida's		
4	About Florida Memory	-"-	De Bry Engravings	https://www.floridam	16th Century		
5	About Florida Memory	-"-	Selected Documents	https://www.floridam	Letters and Other		
6	About Florida Memory	-"-	Florida's Early	https://www.floridam	The Evolution of		
7	About Florida Memory	-"-	Supreme Court	https://www.floridam	Antebellum Cases		
8	About Florida Memory	-"-	1825 Leon County	https://www.floridam	Early Territorial		
9	About Florida Memory	-"-	Spanish Land Grants	https://www.floridam	Maps and Property		
10	About Florida Memory	-"-	1845 Election Returns	https://www.floridam	Voter Lists for		
11	About Florida Memory	-"-	1855 Census Returns:	https://www.floridam	Records from Florida's		
12	About Florida Memory	-"-	Jefferson County	https://www.floridam	Sharecropping		
13	About Florida Memory	-"-	Voter Registration	https://www.floridam	Reconstruction Era		
14	About Florida Memory	-"-	Confederate Pension	https://www.floridam	Soldiers and Widows		
15	About Florida Memory	-"-	Old Confederate	https://www.floridam	Applications for		
16	About Florida Memory	-"-	Fernandina Death &	https://www.floridam	Documenting Death in		
17	About Florida Memory	-"-	Early Auto	https://www.floridam	Records of Florida's		
18	About Florida Memory	-"-	World War I Service	https://www.floridam	Floridians and the		
19	About Florida Memory	-"-	County Guard	https://www.floridam	WWI-era County		
20	About Florida Memory	-"-	WPA Church Records	https://www.floridam	1930s and 1940s		
21	About Florida Memory	-"-	WPA County Histories	https://www.floridam	Local Histories of		
22	About Florida Memory	-"-	WPA Stories	https://www.floridam	Old Florida Folk Tales		
23	About Florida Memory	-"-	Kingsley Papers	https://www.floridam	Slavery in Colonial		
24	About Florida Memory	-"-	Richard Keith Call	https://www.floridam	Collected Papers of		
25	About Florida Memory	-"-	Call and Brevard	https://www.floridam	Prominent		
26	About Florida Memory	-"-	Physician's Journal	https://www.floridam	Medicine on the		
27	About Florida Memory	-"-	Jesup Diary	https://www.floridam	An Account of the		
28	About Florida Memory	-"-	Milton Letterbook	https://www.floridam	Letterbook of		
29	About Florida Memory	-"-	McLeod Diary	https://www.floridam	An Account of the Civil		
30	About Florida Memory	-"-	Gramling Civil War	https://www.floridam	Writings of a		
31	About Florida Memory	-"-	Albert S. Chalker	https://www.floridam	Letters Home from a		
32	About Florida Memory	-"-	Bevins Family Papers	https://www.floridam	Letters from a WWI		
33	About Florida Memory	-"-	Stephens Sisters Jail-	https://www.floridam	Correspondence from		
34						Civil Rights	See All Videos
35						Commerce / Industry	See All Videos
36						Culture	See All Videos
37						Environment	See All Videos
38						Folklife	See All Videos

Figure 8—This ParseHub CSV export shows the information extracted from the sitemap. Each column represents one selector in the sitemap

Figure 8 shows the template from the homepage of Florida Memory, including the about page link and a list of collections present. The first template in a sitemap always starts at column A, row 2, and the next starts after all the data from the first at template is in the file. Row 1 contains the selector names. ParseHub enters any information a selector tells it to extract. It shifts every template after the first down and to the right of the template before it. The first template had selectors that extracted information about the website as well as the name, URL, and description of collections. You can see that the second template starts at column F, row 34. Each column in the exported data corresponds to a selector created somewhere in the sitemap. The blocks of data represent the different templates within ParseHub.

3.3.2. Webscraper.io Export

By comparison, Webscraper.io copies information extracted from selectors on the _root level into each row. The information closest to this root level in the selector graph appears in the leftmost columns, with information further from the root level in the columns following them. Webscraper.io automatically inputs the scraping order into column A and the start URL into column B, which appears in every row in the data. Any other selector in the root level appears in column C, then D, and so on. Once the root level selectors are in place, the selectors beyond them in the selector graph appear in the next columns. Figure 9 shows the copying of information into each row, as well as the information appending itself to the columns to the right.



	A	B	C	D	E	F	G	H	I	J	K	L
1	web-scraper-order	web-scraper-start-url	creativework	about	about-href	contributor	project_title	project_title-href	creativework_rights	organization-src	PrincipalInvestigatorContact	PrincipalInvestigatorContact-href
2	1536270053-104	http://coloredconventions.org/	[{"creativework":"ABOUT THE COLORED CONVENTIONS\nFrom 1830 until the 1890s, already free and				Bishop Henry McNeal Turner	http://coloredconventions.org/exhibits/show/bishophmturner	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License		Contact Us	mailto:coloredconventions@udel.edu
3	1536270056-163	http://coloredconventions.org/	[{"creativework":"ABOUT THE COLORED CONVENTIONS\nFrom 1830 until the 1890s, already free and once captive Black people came together in state and national	About Us	http://coloredconventions.org/about-us	CALEB OWENS is an undergraduate student pursuing majors in History, English, and Philosophy. Since joining the Colored Conventions Project, Caleb works with the			Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License		Contact Us	mailto:coloredconventions@udel.edu
4	1536270056-165	http://coloredconventions.org/	[{"creativework":"ABOUT THE COLORED CONVENTIONS\nFrom 1830 until the 1890s, already free and	About Us	http://coloredconventions.org/about-us	MARISSA ROSS is a sophomore at the University of Delaware studying English with a concentration in			Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License		Contact Us	mailto:coloredconventions@udel.edu
5	1536270053-125	http://coloredconventions.org/	[{"creativework":"ABOUT THE COLORED CONVENTIONS\nFrom 1830 until				#CCPclass on Twitter	http://coloredconventions.org/cc-p-class	Creative Commons Attribution-NonCommercial-ShareAlike 4.0		Contact Us	mailto:coloredconventions@udel.edu

Figure 9—Colored Conventions Webscraper.io export

The export for Colored Conventions seen in Figure 9 shows the information input into the first three columns: the scraping order, the homepage, and content information from the homepage. Subsequent columns represent selectors within different tiers in the selector graph, from the about page in columns D and E, to contact information in columns K and L. In both ParseHub and Webscraper.io, the exports provide the information users need in order to complete their larger projects.

3.4. Comparing the Scrapers

The table below compares the key components of each scraper. The entries in bold and italics indicate which scraper is better for that component.

ParseHub	Webscraper.io
5 total active projects (public on servers)	***Unlimited active projects (private)***
Free option	***Free option***
Moderate learning curve	Moderate learning curve
Friendly GUI	Confusing GUI

Uses HTML for selectors	***Uses HTML and CSS for selectors***
Links scraped as name only	***Links scraped as name and URL***
Exports data as CSV and JSON	Exports data as CSV

The biggest and most pressing difference is that Webscraper.io does not limit the number of projects. ParseHub has a maximum of 5 projects at one time, all of which are publicly viewable by other ParseHub users. Both have free options. Each takes a little bit of time to learn. Figuring out what the various fields and buttons do can be difficult at times, but the learning curve between them is about equal. ParseHub is more visually appealing, with a clear method of organizing the different levels of selectors, while the sitemaps in Webscraper.io can appear dense. The main difference is that the selection and exporting of data is more powerful in Webscraper.io. It uses both HTML and CSS, while ParseHub only uses HTML. Webscraper.io also automatically extracts the name and URL for all link selectors, as opposed to the name only.

Each scraper provides a way to see the scraped data before exporting it as a file. However, ParseHub users can only see data from the specific selector being viewed, while Webscraper.io also has a dedicated section that previews all the data in a scrape. Both export CSV files, but ParseHub can also provide separate JSON files. However, the lack of a direct JSON export is not so terrible, as CSV files are more interoperable.

Either web scraper can be useful depending on the project at hand. ParseHub is a good choice to learn what web scraping is and how sitemaps and selectors work, but it does not work for projects that need to scrape a large number of projects. By contrast, Webscraper.io is a more robust scraper that should be able to handle most projects but can get a little confusing in terms of the structure of the sitemaps, which is where the selector graph becomes useful.

Webscraper.io became our primary scraper from this point forward because it had the most flexibility in scraping projects and did not have as many drawbacks as ParseHub. There is no limit to the projects it can

test, and the export files it produced were much more comprehensive than those produced by ParseHub. It was at this point in the project that we needed a standard list of terms. This necessitated that we create a metadata dictionary to keep track of the various terms used in web scraping and to define the terminology.

4. Metadata Dictionary

Each sitemap up to this point had tended to use the vocabulary within the project itself, so we needed to create one standardized list of terms to use throughout all the sitemaps: a metadata dictionary. This involved comparing the SHARE metadata schema with the terminology used in the test projects and the Dublin Core utilized by Omeka to see if there was any overlap between terms, as they needed to remain as consistent as possible throughout the project. This process also highlighted those terms that did not appear in either the SHARE schema or in the sitemaps already created. The terms needed to use the forms found in the SHARE schema,[13] meaning a term like 'Project Title' became 'CreativeWork.' The terms in each sitemap start with a title and URL for the project itself and then become more descriptive. We entered titles and descriptions for every smaller project into the dictionary, including those for items, collections, and exhibits. We also needed to define terms for rights information and contributors and organizations. We entered the list of terms into a spreadsheet, where we compared them to the SHARE metadata schema, the terms used in the web scrapers, and the Dublin Core metadata embedded into Omeka. We labeled any that were common with the term from the SHARE schema, which tended to be project titles, descriptions, and contributors to the project. There were some Dublin Core terms that had no direct translation in SHARE, such as titles and descriptions held within each item. These kept the original terms in Dublin Core, but we labeled them with 'item_' to denote that they were item terms. For example, we used the label 'item_title' for all item titles, 'item_description,' and so on.

To build the dictionary, we created a number of different sections.[14] Figure 10 highlights the different types of information used to help define the metadata terms. First, we listed each term used in the web scraper. Then, we compared each term to the list of terminology in the SHARE schema. We matched the SHARE term to the scraper term that most closely matched it. This also helped us keep track of which SHARE terms were used in the scraper. The terminology stayed the same in both SHARE and the scraper whenever possible. Each term in the dictionary includes a description that defines that particular term, which helps users differentiate between the title of an overall project, which SHARE calls 'CreativeWork,' and smaller projects, like collections and exhibits posted to the larger project. We also included where the term falls in the organizational hierarchy, such as top-level information, or whether it refers to a collection or an exhibit. Top-level information includes things like information regarding a website as a whole, such as its name and description.

	A	B	C	D	E	
1	SHARE Metadata Element	Webscrapper Metadata Label	Level of organization (Collection, Exhibit, Item)	Description	Example URL	
2	creativework	creativework	Top Level	The name of the project or website scraped	http://newdeal.oucreate.com	Making Modern America
3		creativework-href	Top Level	URL address for the creativework	http://newdeal.oucreate.com	
4	creativework_description	about	Top Level	A description of the creativework	http://newdeal.oucreate.com/about	This site was created by students at the University of Oklahoma New Deal during Fall 2015. During the course, students became trips. Through weekly workshops, students gained the skills and
5	project_title	project_title	Collection, Exhibit	Title, list, and/or links to exhibits or collections held within the creativework	http://www.grandeguerra.unito.it/collections/show/1	Archivi
6	project_description	project_description	Collection, Exhibit	The description of a collection or exhibit.	http://www.grandeguerra.unito.it/collections/show/1	http://www.grandeguerra.unito.it/collections/show/1
7		project-href	Collection, Exhibit	The URL for a collection or exhibit	http://www.grandeguerra.unito.it/collections/show/1	I volti e il ricordo degli studenti caduti al fronte; i nomi degli stuc l'assistenza ai soldati al fronte e ai profughi; la mobilitazione dell
8	contributor	contributor	All	People who contributed to the creation of the project	http://coloredconventions.org/about-us	Founding Faculty Director and Co-Founder: P. GABRIELLE FOR history and culture...
9	description	description	All	Text that describes an exhibit, collection, or project, usually topically	The University of Oklahoma	This exhibit profiles buildings and artwork on the University of Ok still exist today, some have changed used, and others no longer more projects.
10	publisher	publisher	All	Entity responsible for posting or distributing a creativework, project, collection, or item	http://coloredconventions.org/items/show/1468	New York Public Library
11	creativework_rights; project_rights	rights	All	Any rights the creativework, project, or exhibit is held under		Creative Commons Attribution
12	PrincipalInvestigatorContact	contact	Top Level	Any contact information for the project, primary investigator, or contributor.	http://coloredconventions.org/	coloredconventions@udel.edu
13	Organization	organization	Top Level	A group that contributes information or support to the creation of the project	http://coloredconventions.org/	University of Delaware
14		item_title	Item	The title for an item	Oklahoma History 1930s: Reconstructing Memory	Oklahoma History 1930s: Reconstructing Memory
15		item_description	Item	A description of the item	Oklahoma History 1930s: Reconstructing Memory	A lesson plan created by Donna Moore and Dalton Savage
16		item_creator	Item	Person or entity responsible for creating the item	Oklahoma History 1930s: Reconstructing Memory	Moore, Donna
17		item_date	Item	Date the item was created OR date the object being described was created	Oklahoma History 1930s: Reconstructing Memory	2016-05-13
18		item_rights	Item	Any rights the item is held under	Oklahoma History 1930s: Reconstructing Memory	CC BY-NC-SA 2.0
19		item_tags	Item	A label attached to an item to group common items together, usually separated by a comma	Oklahoma History 1930s: Reconstructing Memory	Education, Lesson Plans

Figure 10—The metadata dictionary has the following columns from left to right: the SHARE term, label in the scraper, hierarchical tier, description, and an example

The first column in Figure 10 is the list of terms pulled from the SHARE schema. The blank cells toward the bottom correspond to the Dublin Core terms that did not have direct corollaries. The second column is the list of terms that are used within the web scrapers. These must follow a rules to avoid errors: they must be three or more characters and not contain a space, period, or dollar sign. The third

column lists the level of hierarchy for each term, such as whether it refers to an exhibit or collection or top-level information such as the website title. The description column contains the definition for each term. The example column includes examples pulled directly from the web scrapers or export files.

The examples for each term were pulled from test scrapes in order to help understand them. Using Colored Conventions as an example, the 'CreativeWork' term is Colored Conventions while the information under 'About the Colored Conventions' falls under 'CreativeWork_ description.' We decided that the 'Project' label covers both exhibits and collections because both are established by the creators of the project using items already created.

Once we finalized the metadata dictionary, we updated any already created sitemaps with SHARE vocabulary within Webscraper.io. The scraper allows users to edit any part of a selector they created by clicking on the 'edit' button after which they can change the fields that need updating without having to recreate the entire selector (See Appendix A §4 Creating a Selector for more information). This is helpful when a selector needs to be renamed or moved between parent selectors. Once we made these changes, we added any necessary, additional information to the sitemaps. This was typically rights information and organizations. Organizations also needed some clarification, because there was concern over whether they also belonged under the contributor term. For this project, we limited contributors to individual persons, while organizations referred to groups or large entities.

5. Foreign Language Websites

One question arose while creating the metadata dictionary: What if an Omeka project is not in English? We didn't know how the web scrapers would handle non-English sites. We used the translation feature built into Google Chrome to provide rough translations of websites that were not in the default language of the browser. This meant Webscraper.io could handle scraping a foreign-language website.

Users can set Chrome to translate foreign websites into English to create selectors for the information they want to extract. Then they can scrape websites without any knowledge of the original language. Websites are set up exactly the same regardless of language due to their use of HTML. The original language is retained when the data is extracted as long as Chrome is not set to automatically translate foreign languages. Figure 11 shows that any text will stay in the language entered, such as sitemap titles and selector names.

	A	B	C	D
1	web-scraper-order	web-scraper-start-url	creativework	about
2	1534367917-10	http://bmn-renaissance.nancy.fr/	BIBLIOTHÈQUE RENAISSANCE À NANCY	Réalisé à partir d'une sélection d'oeuvres lorraines numérisées par la Bibliothèque-Médiathèque de Nancy, le site appelle à la découverte grâce à différents parcours dans l'exposition et la galerie des graveurs lorrains. On peut aussi se laisser prendre au jeu avec les oeuvres : memory et puzzles, accès par thèmes, carte et même une baleine ! Les Bibliothèques universitaires de Nancy fêtent aussi la Renaissance au travers d'une exposition dont vous pourrez retrouver le contenu ici.

Figure 11—Foreign language website export

As with any sitemap, the selector names will become the column headers in the exported CSV file. This means that while these headers will remain in the language typed into the web scraper, he extracted information will remain in the original language, so the selector labels are very important for these sites. This is because the

user manually enables the translation feature, which Webscraper.io does not do during a scrape. The selector names may become the only understandable part of the export file, meaning it is possible to not recognize to which category the information belongs if labels are incorrect or incomplete. Foreign language websites are therefore not necessarily an impediment to scraping websites. Using Chrome's translation feature will allow users to successfully extract information from them.

6. Project Updates

At this point in the project, we decided to focus on scraping only the upper levels of information, which includes information regarding the website as a whole, such as its name and description, as well as organizations and contributors. We extracted collection and exhibit names and descriptions along with their links, which captured main topics and general information that the projects contained without requiring the set-up of complete sitemaps for entire websites. We left out contents and items to keep the amount of information manageable. This helped us avoid the duplications of metadata terms that exist within individual items and the larger project information. Since items have their own sets of information, they need their own selectors within the web scrapers, which includes labels that become column headers in the export file. If these labels are not carefully considered, it can be confusion which title belongs to an item, a collection, or an exhibit when only viewing the exported CSV file.

7. General Scraping Problems

One SHARE metadata label that became a larger issue was the 'Project' element. Both collections and exhibits fall under this category, but Omeka separates them onto different pages with their own structures. For example, the creators of Florida Memory categorized its collection into things like photographs and videos. There are links on the Florida Memory pages that direct to exhibits, collections, or items. These links lead to different pages, usually with their own div and container

names in the HTML code. This often presents complications for the web scrapers. As seen in Figure 12, there are six thumbnails on the homepage that link to pages that each have multiple projects.

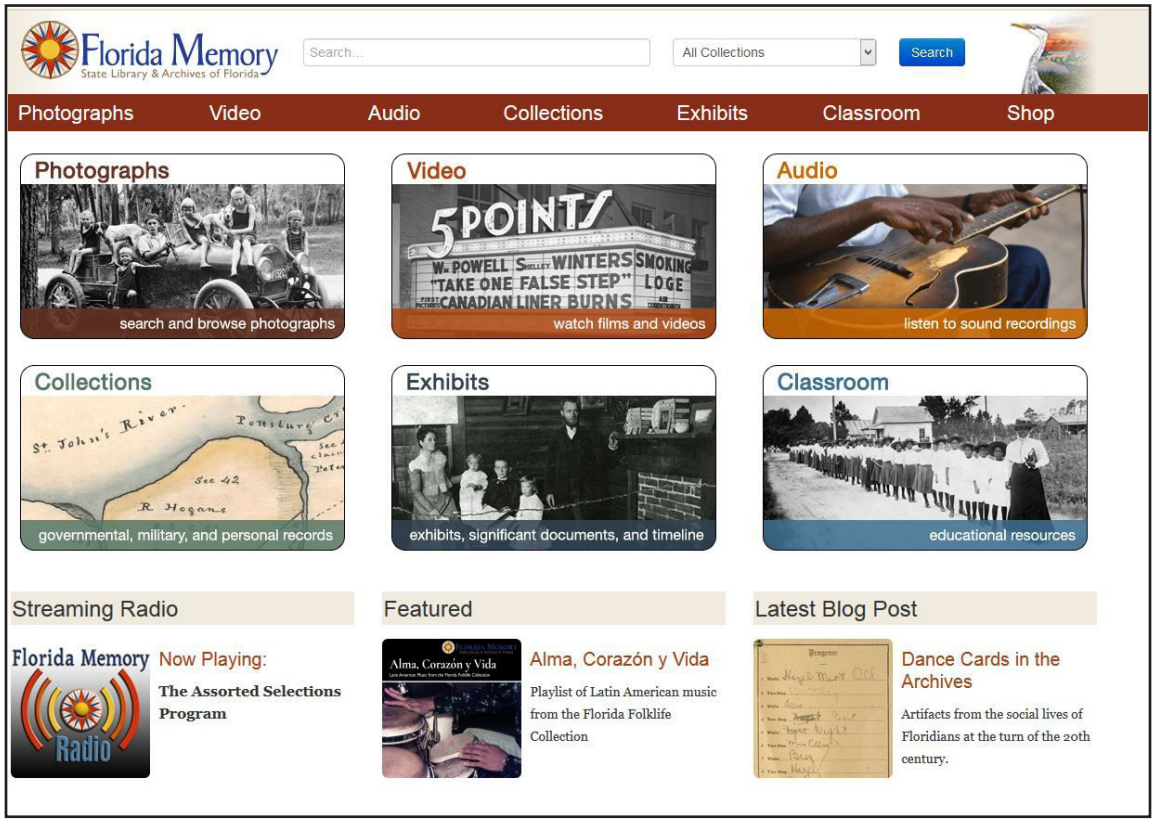


Figure 12—Florida Memory Homepage. Note the main thumbnails with different media types

This wreaks havoc with the SHARE metadata labels because the exhibits and collections are split up on so many different pages. Our solution for the SHARE project was to create a selector with its own unique names for each page and then use data cleaning software like OpenRefine to collapse this metadata into one column in the exported CSV file. In Figure 13, you can see how complex the sitemap became in the selector graph in Figure 13.

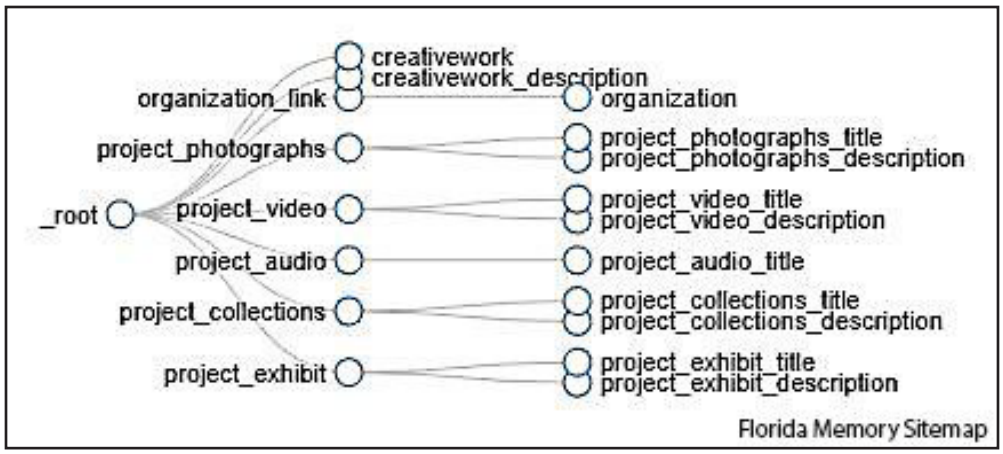


Figure 13—Florida Memory sitemap

Webscraper.io can select multiple links in one type of HTML code, such as a navigation menu, by using CSS nth-of-type coding.[15] This code sets which item in the navigation menu is included. Users can set the selector for all links before or after a specified link or choose only certain links that may not be next to each other in the menu. The code goes in the section labeled 'Selector,' where the HTML code is input after clicking on an element.

An issue arose while extracting organizational information. Several websites had the organization posted as an image rather than text. Images can be scraped, but Webscraper.io only extracts HTML information, which may or may not contain the information users require. It depends on how the creators entered an image into a project, as any text within the image itself cannot be extracted. This is usually a problem with organization logos, and there is also the rare instance of a project title acting as a link to the homepage. The problem with this is that the image selector will not work as intended, but the link selector will. Following best practices, using alternate image titles can help future researchers avoid this problem.

8. Conclusion

Our intention with this project was to help create a database covering a wide range of Digital Humanities projects. The database does not require every piece of information presented on these website, only

the information that helps outsiders understand the topics held within them. Web scrapers can help extract information from any project published on the Web, not just within Digital Humanities. Any tool has some sort of a learning curve, but the information they provide can help with discovering and sharing information that exists on the Web. Web scrapers could be used in projects similar to the SHARE project to collect information regarding a theme or type. The decision in this project to only scrape higher levels of information need not apply to every project. The next steps for this project in particular are to see how this process can apply to non-Omeka projects, such as those in Drupal or Wordpress. The extracted information from these different methods could then be analyzed to see how different methodologies exist within the overall DH community. It remains to be seen how the exported data interacts with SHARE and its functionality. Another possible direction is to see if using information from items and collections, given copyright permissions, could be useful in different types of analyses.

Appendix A

Webscraper.io
A How-to Guide for Scraping DH Projects

1. Introduction to Webscraper.io

Webscraper.io is a free extension for the Google Chrome web browser with which users can extract information from any public website using HTML and CSS and export the data as a Comma Separated Value (CSV) file, which can be opened in spreadsheet processing software like Excel or Google Sheets. The scraper uses the developer tools menu built into Chrome (Chrome DevTools) to select the different elements of a website, including links, tables, and the HTML code itself. With developer tools users can look at a web page to see the code that generates everything that is seen on the page, from text, to images, to the layout. Webscraper.io uses this code either to extract information or navigate throughout the overall page. This is helpful for users who don't have another way to extract important information from websites. It is important to be aware of any copyright information listed on a website. Consult a copyright lawyer or professional to ensure the information can legally be scraped and shared.

This guide was developed alongside a project for extracting information from websites using content management systems like Omeka to make them discoverable. We focused on Omeka because it has been widely adopted by the Digital Humanities community and provides a range of information presentation types, including titles and descriptions of each project as well as the collections and exhibits held within them. It also extracts any contributors and organizations listed. The screenshots used in this documentation are pulled from Colored Conventions, which is held under a Creative Commons Attribution Non-Commercial Share-Alike 4.0 International Copyright license. Colored Conventions is a straightforward website with little complex coding, which allows Webscraper.io to function at its best.

1.1. Installing Webscraper.io

Type 'webscraper.io' into your URL bar to navigate to the scraper's website. The site has a wealth of documentation as well as a very active forum. Webscraper.io regularly updates both of these sections with information that can help resolve specific issues that arise. To install

the extension itself, click on the blue 'Download Free on Chrome Store' button.

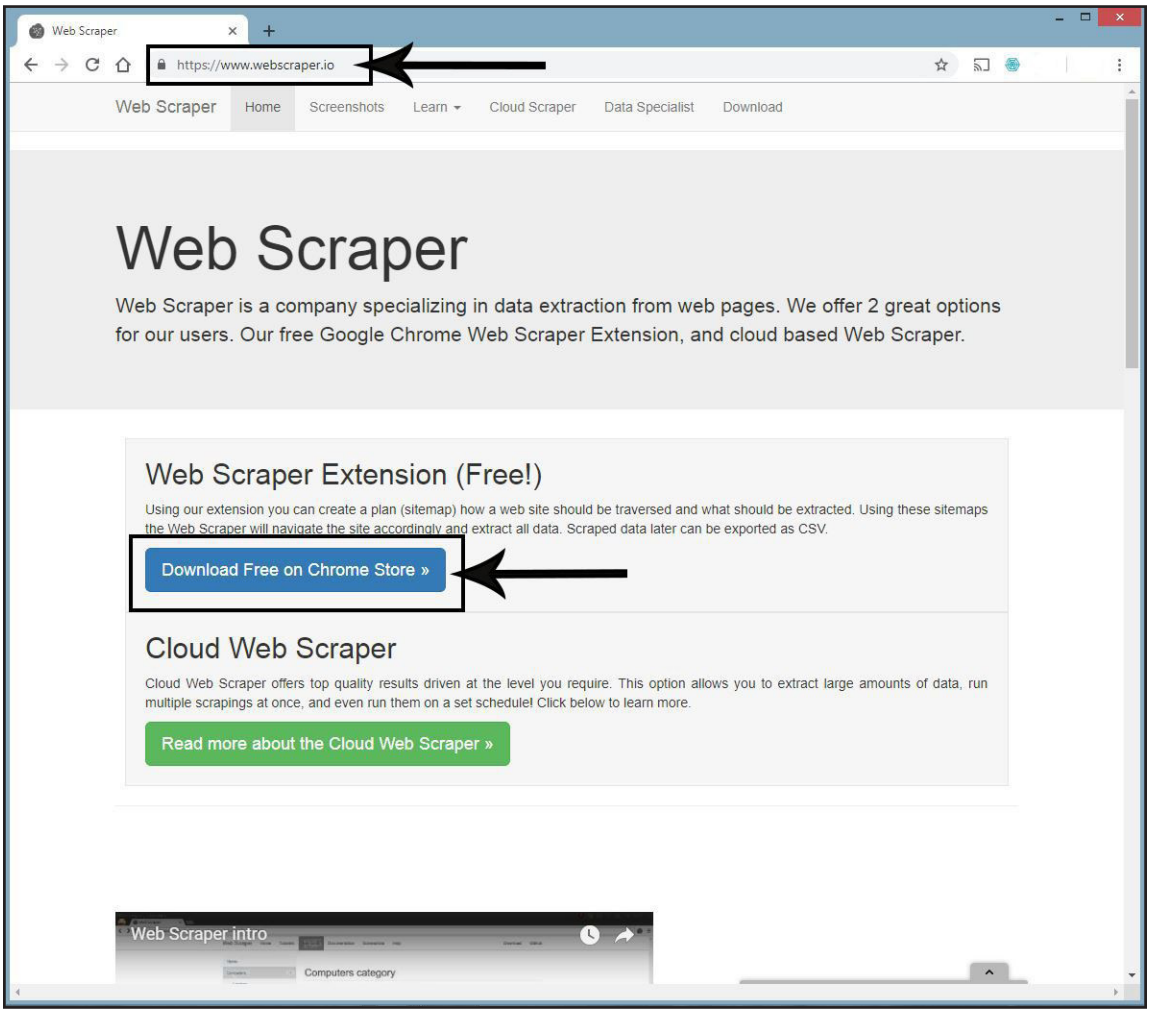


Figure 1—Webscraper.io homepage

Click on the green button that says 'Add to Chrome' at the top right corner of this new page tol install the extension. If the installation is successful, the button will gray out and read 'Added to Chrome.'

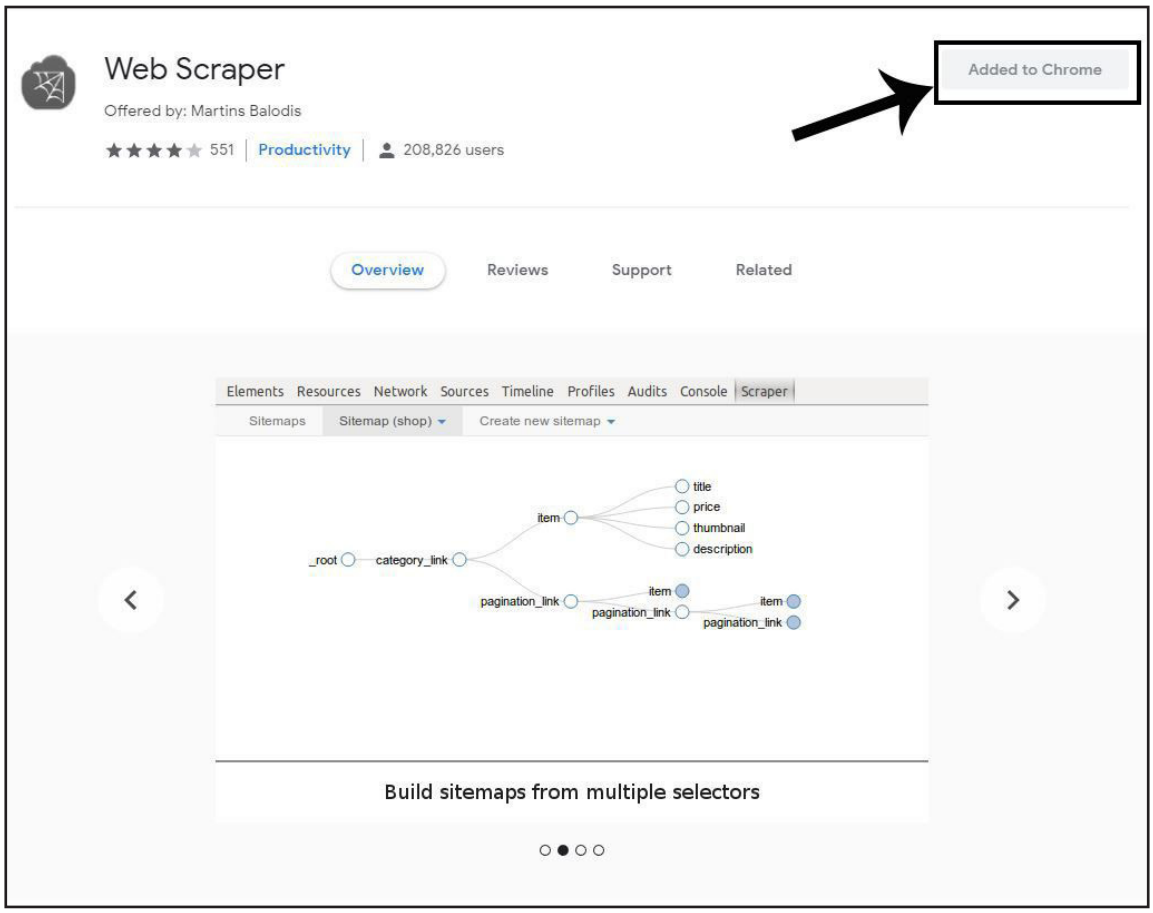


Figure 2—Click the 'Added to Chrome' button to install Webscraper.io

1.2. Navigating to Webscraper.io

The scraper can be found in the Developer tools menu. The first way to get to this menu is to press the F12 button on either a Mac or PC. The other way is to click on the three vertical dots in the upper right corner of the window. Both these ways bring up the browser menu, which is the same menu that opens a new tab or window as well as the history or print panels. Hover the mouse over the 'More tools' text roughly two-thirds of the way down to show a sub-menu. This sub-menu has even more options, but the web scraper is in 'Developer tools' at the bottom of this new menu. Click 'Developer tools' to open the panel. See Figure 3 below.

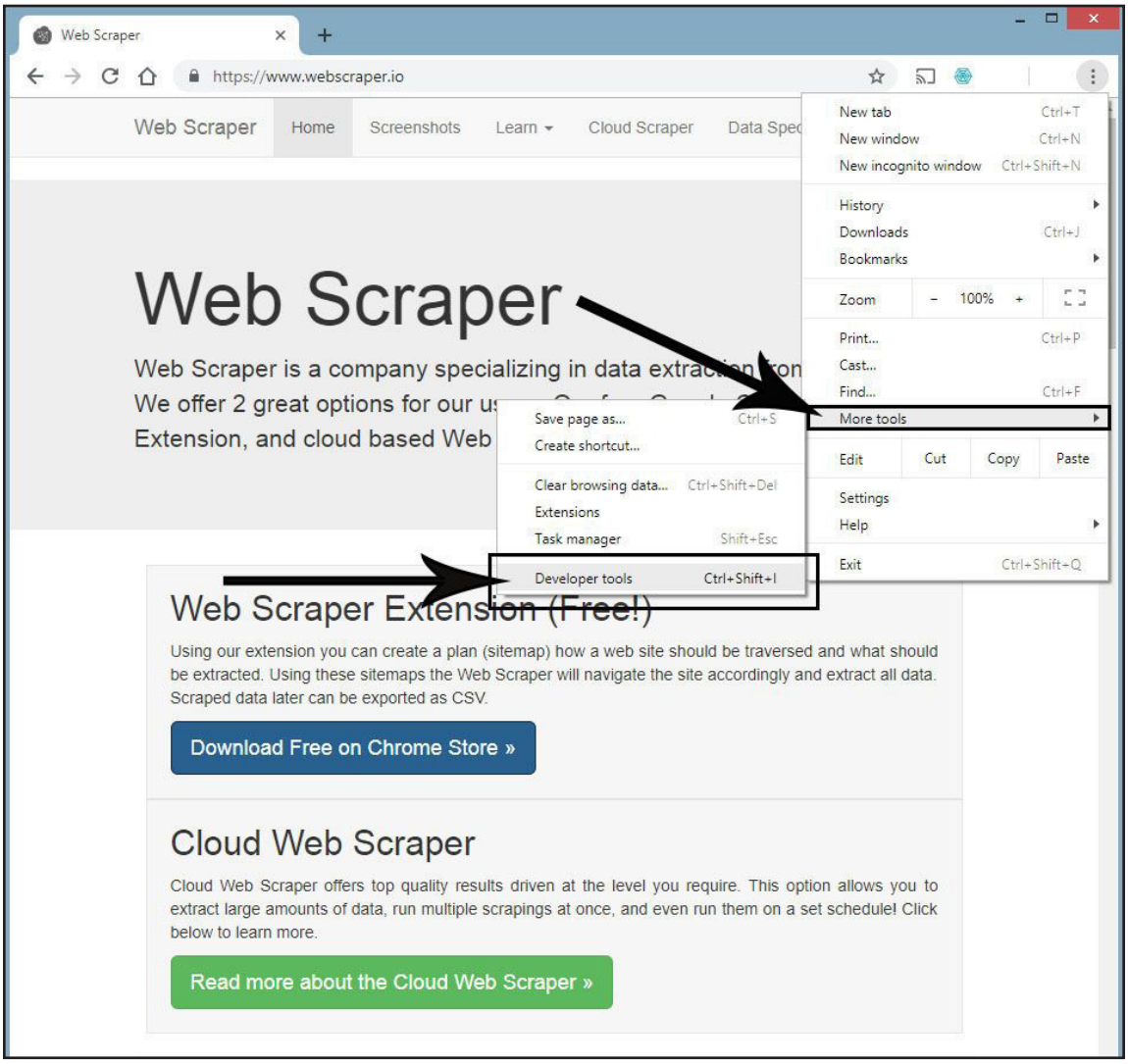


Figure 3—Opening 'Developer tools'

This opens a new panel at the bottom of the browser. Webscraper.io is the last option in this panel, as you can see in Figure 4 below.

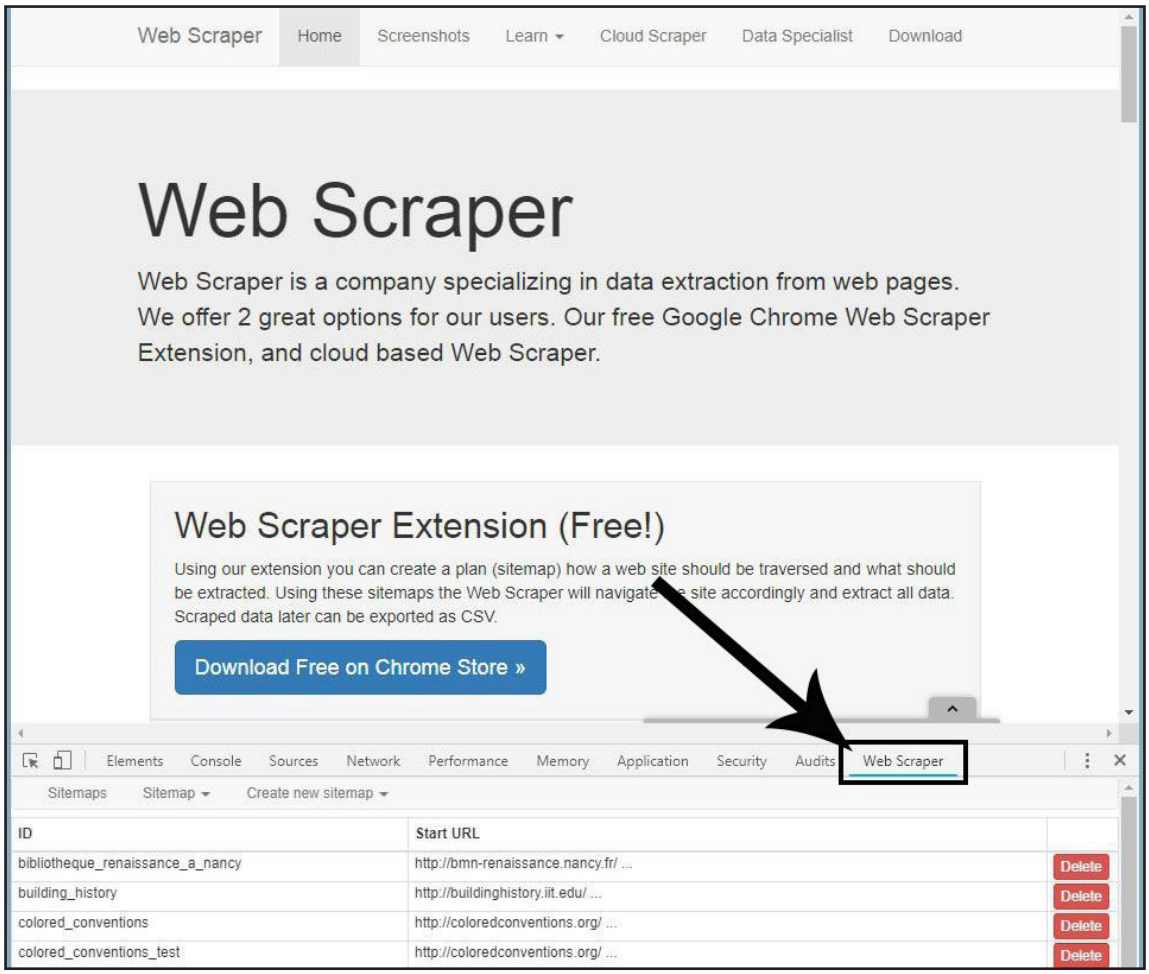


Figure 4—'Web Scraper' panel

The first window that appears when navigating to Webscraper.io is the Sitemap panel (See Creating a Sitemap). A sitemap organizes all the information required for scraping a particular website. It will be blank at install, but once you create sitemaps, they will appear here. The first column lists the ID, or name, of each sitemap. The second column is the URL or web address for the first page of that sitemap.

2. Creating a Sitemap

2.1. Sitemap Menu

Webscraper.io automatically opens to the Sitemap Menu, which lists all of the user-created sitemaps in the scraper. Here, users can see all of their sitemaps alongside each starting URL. They also have the option

to delete sitemaps. Be careful not to delete sitemaps, as they cannot be recovered unless there are exports of them saved elsewhere (See Exporting Sitemaps). Click a sitemap's title or URL to open it.

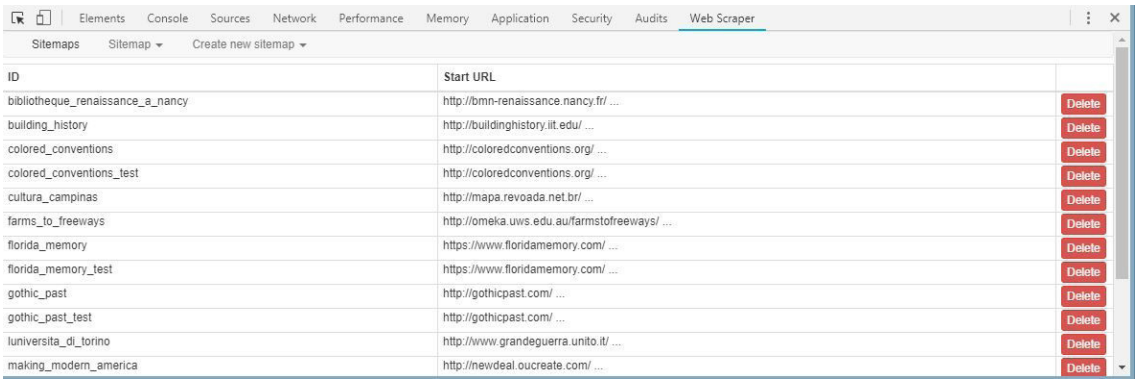


Figure 5—Sitemap menu

Sitemaps serve to organize all the information about scraping a particular website in one location. They house the various selectors (See Creating a Selector) and instruct the web scraper what the titles and starting URLs are for them to scrape. To create a sitemap, click on the 'Create new Sitemap' button. Then you can either import a previously built sitemap or create a blank sitemap.

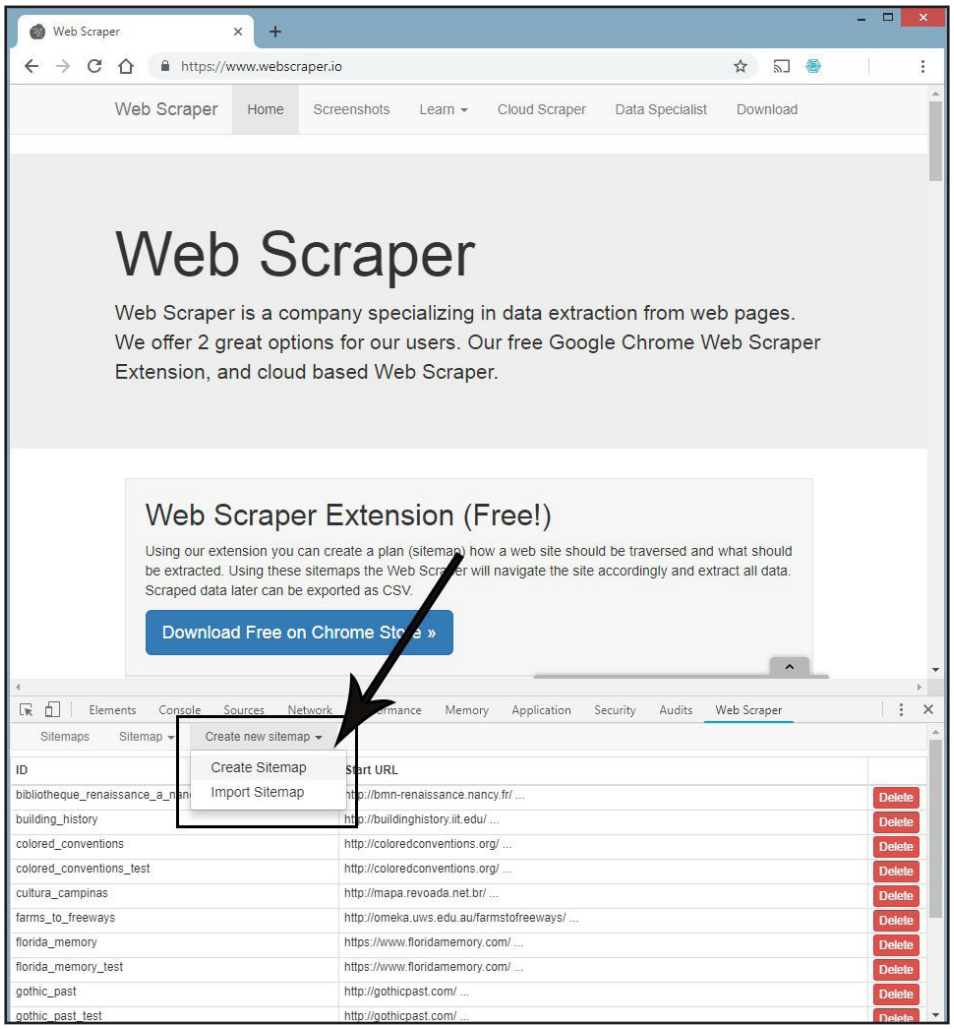


Figure 6—'Create Sitemap'

2.2. Importing a Sitemap

A user who has already created a Webscraper.io sitemap has the option to export and share that sitemap with other users to import in their own web scrapers (See Exporting Sitemaps). The 'Import Sitemap' button creates a sitemap, which can then be manipulated. Importing a sitemap requires the JavaScript Object Notation (JSON) that another user's instance of Webscraper.io generated. Clicking on the 'Import sitemap' button brings up two text entry fields. The user copies and pastes the JSON, which is formatted in a particular way, into the larger of the two fields.The user can rename the sitemap something distinct from the imported JSON code in the second box to ensure that there is

no duplicate sitemap within Webscraper.io. For group projects, it also helps to keep track of the date you imported, a sitemap, or who worked on it, by adding the relevant information to the end of the title.

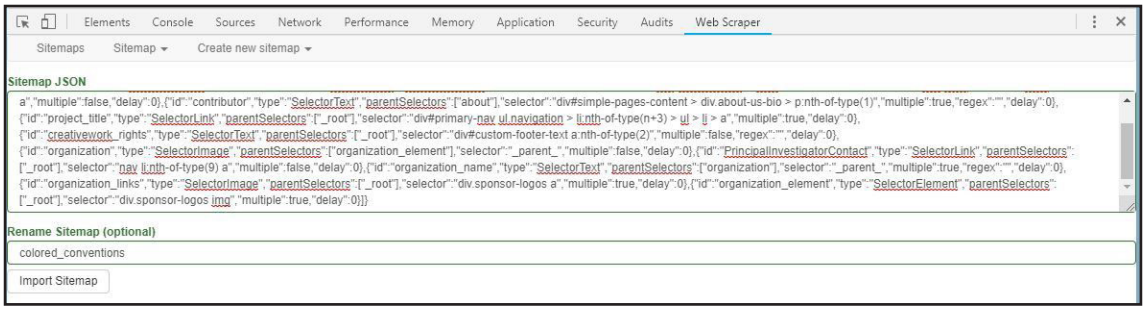


Figure 7—Import a sitemap

2.3. Creating a Blank Sitemap

The 'Create Sitemap' button opens a window similar to a window opened by the 'Import Sitemap' button. The difference here is that there is no previous information, and the new sitemap will have no scraping information within it. The user creates a new sitemap at the beginning of any project in order to create the selectors that will extract information from a website. This requires the sitemap name and the URL for a website, which is usually the homepage. The sitemap title has a few rules: it cannot have any capital letters, limits the special characters it recognizes, and must start with a letter. It may be helpful to copy and paste the URL into the 'Start URL' field to avoid errors.

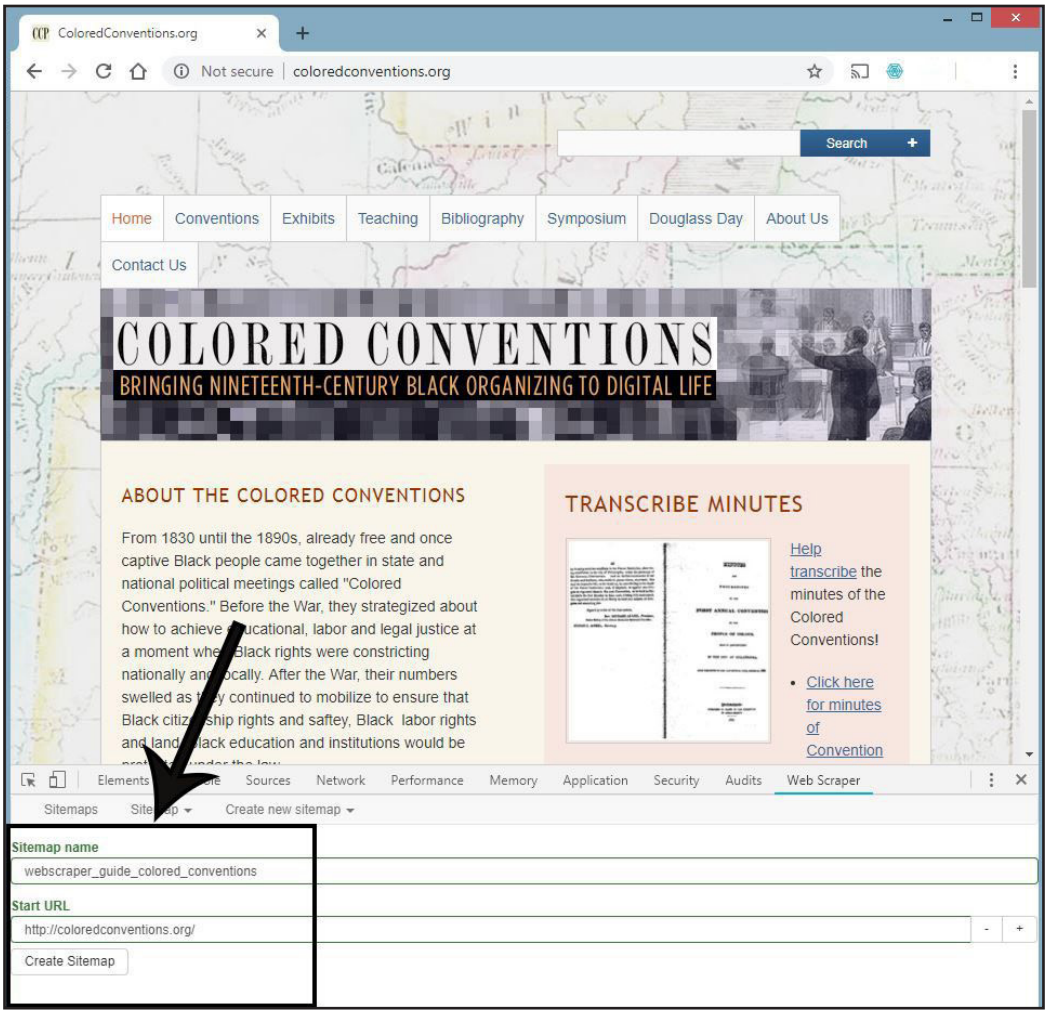


Figure 8—Creating a blank sitemap

Once the user enters the title and URL, he or she should click on the 'Create Sitemap' button to add it to the web scraper.

2.4. Editing Project Metadata

If the sitemap name or start URL ever need to be changed, users can do so in the 'Edit metadata' panel in cases where there are errors or if the project belongs to a larger project outside of the web scraper that mandates a change. The sitemap name is almost always the information that needs to be changed, not the start URL.

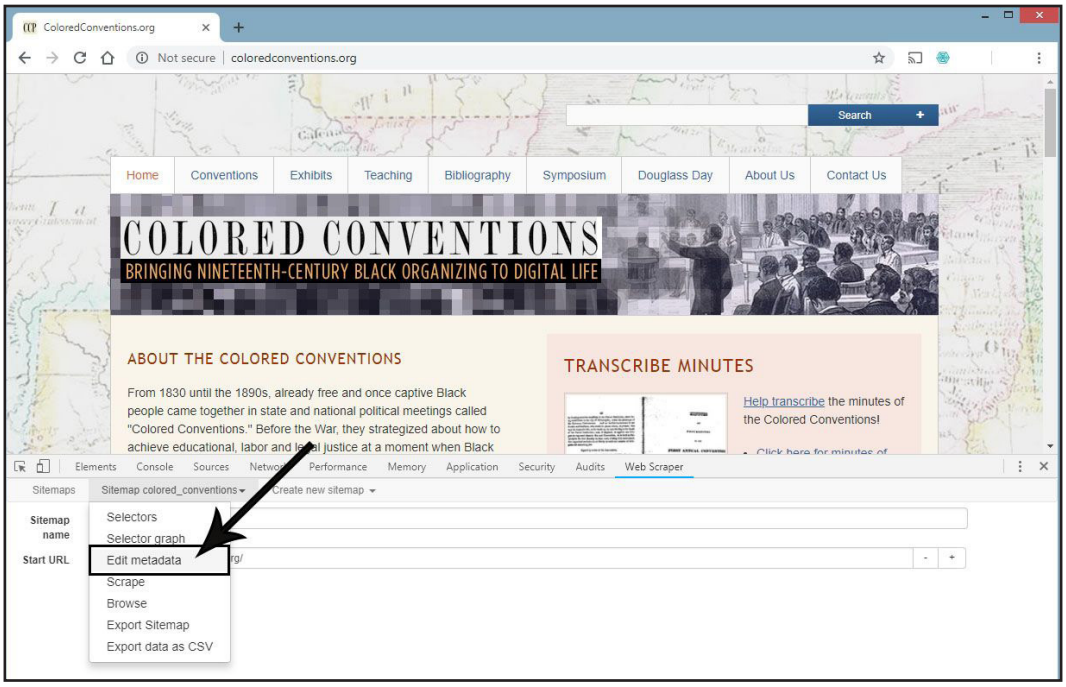


Figure 9—'Edit metadata' panel

Fields are changed in the same way as when creating the sitemap. Be aware that changing the start URL can affect prebuilt selectors in unintended ways, especially those that select unique information. Since the selectors only use the HTML, any selector on the homepage will look for that code. If the homepage changes, the selectors will look for code that may not exist on the new page and then return a 'null' in the scraped data. Another issue that may arise is that the scraper may extract the wrong information. The HTML at the start URL may not change, but its contents may have. This can lead to confusion when reviewing the scraped data. It is wise to double-check selectors so that they still act as expected after changing the start URL.

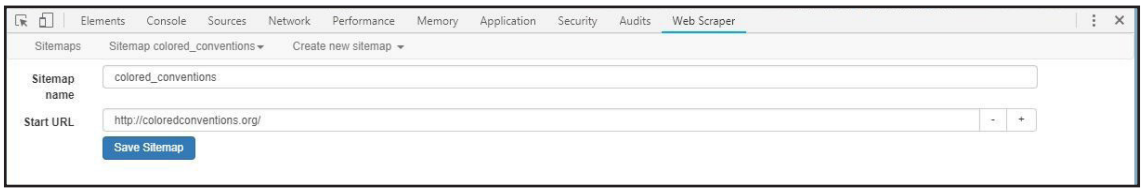


Figure 10—Metadata fields

3. Selector Graph

The selector graph is a visual aid that shows the sitemap hierarchy of selectors, including which selectors are linked to the _root (homepage) and then which selectors are attached to the homepage. This repeats until all selectors are visible, helping users understand where the various selectors are in relation to others. As you can see in Figure 11, the homepage has a number of selectors attached to it, including project_title, PrincipalInvestigatorContact, and about. The selectors beyond the first level, like the about selector, show that there is information to scrape beyond what is selected in the first selector.

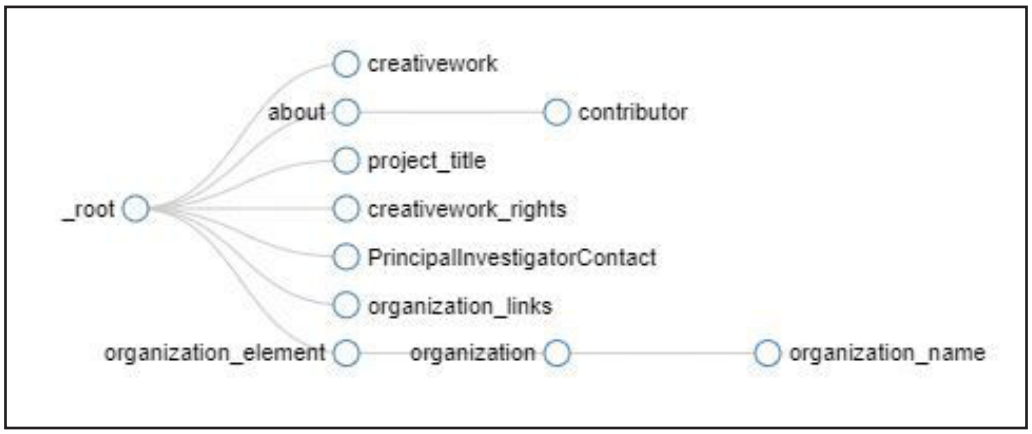


Figure 11—Selector graph

4. Creating a Selector

Users build sitemaps with selectors. These selectors tell Webscraper.io what to do with each element on the website, including extracting a paragraph of text, clicking on a link, or scrolling down the page. Selectors are essentially the instruction manual for the web scraper. Webscraper.io will only do what the selectors within a sitemap tell it to do.

Users can input information into a number of different fields for each selector, including ID, Type, Selector, Multiple Checkbox, Regex, Delay, and Parent selectors (See Figure 11). The ID, short for Identifier, is the title or label for the selector. It will appear in the selector menu and the selector graph. Keep in mind that it will also become the column

header in the exported CSV file. The most common IDs for DH projects are titles and descriptions.

The type menu tells Webscraper.io what kind of information is being used and how it will manipulate it. This helps the scraper organize data or navigate the website. The most common types are text, which will be extracted, and link, which will extract both the link text and URL. The link selector also tells Webscraper.io that there is information on the linked page, which allows users to add selectors to new pages for scraping. Other types that we did not test are HTML selectors and the element selection.

The selector field is the most important field other than the ID. It is where users select the elements on the web page that they want to use.

Once the sitemap is open, click on the 'Add new selector' button to start building the sitemap (See Figure 12). This will open the Selector panel (See Figure 13).

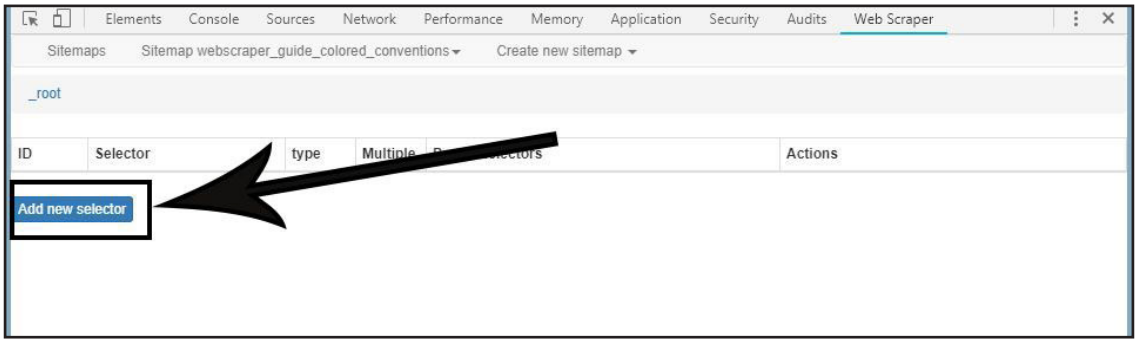


Figure 12—'Add new selector'

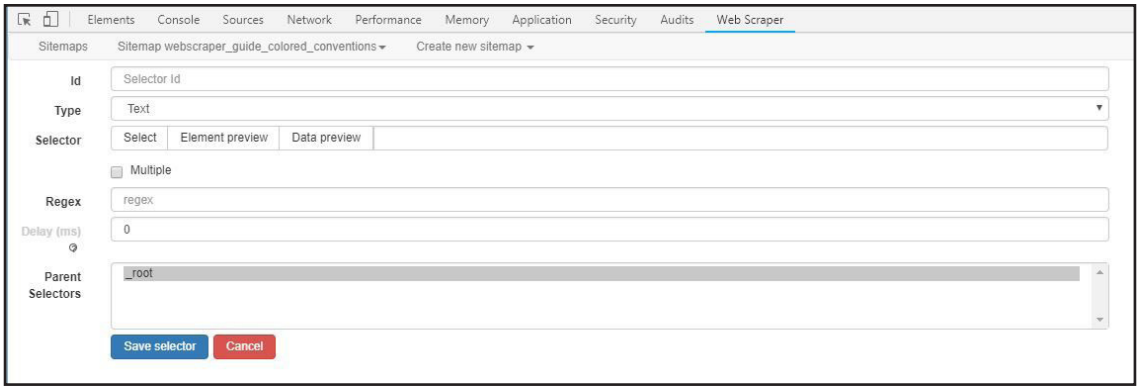


Figure 13—New selector panel

Type in the name of the selector in the 'Selector ID' text box. There are no rules regarding the ID. If the web scraping is part of a larger project outside of the scraping itself, we suggest using the vocabulary set up for that larger project.

Clicking on the 'Select' button makes a green highlight appear around the different HTML elements on the web page. It highlights both HTML tags, like h or p tags, and CSS div and container tags. Clicking on an element adds the code to the selector bar. Each green highlight turns red when you select it. The checkbox in the select bar allows users to select multiple types of tags, which is useful for keeping a title with its description. The 'Done selecting' button adds the selections to the selector text box. Users can also edit the text, if necessary.

The 'Element preview' button puts a red highlight around all the elements in the selector code. This helps ensure that all the elements are selected. By contrast, the 'Data preview' button opens a pop-up window with a snapshot of the data that is set for extraction within this selector when Webscraper.io scrapes the sitemap.

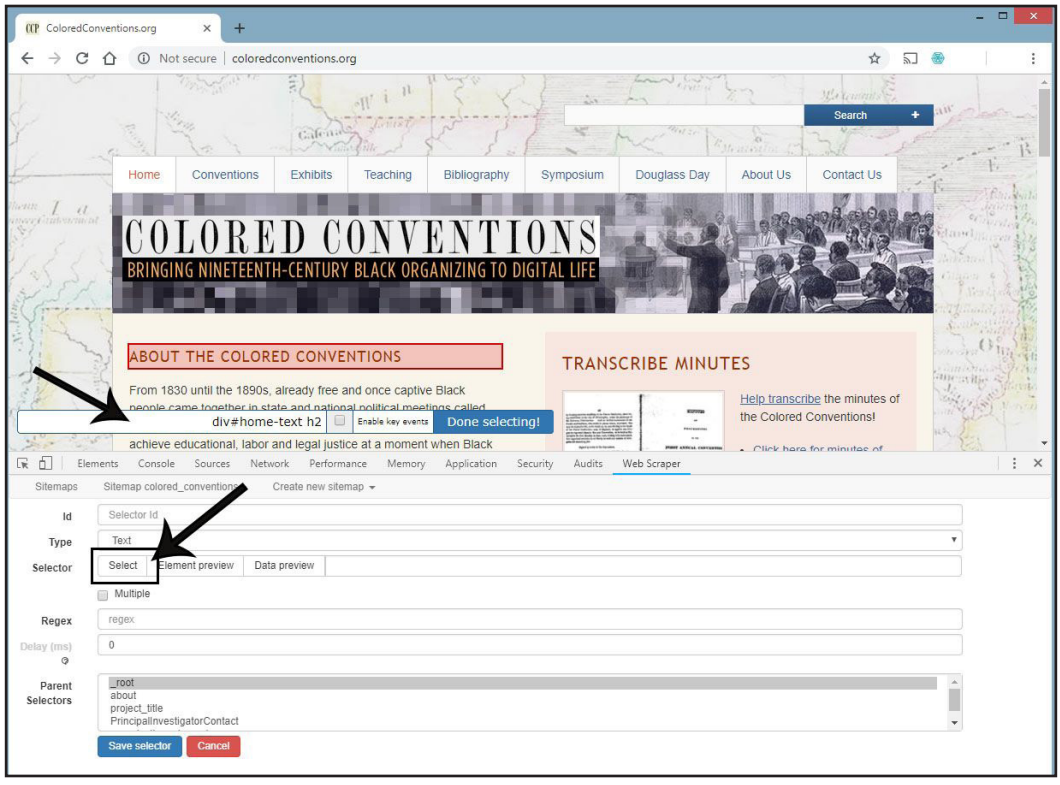


Figure 14—Select HTML element

The Multiple checkbox tells Webscraper.io to extract more than one of the selected elements. This is helpful when there are lists or navigation links with more than one of the same tags on the page.

We did not test the Regex box during this project, which enters regular expressions into the export file to manipulate data. We also did not test the delay, which tells the scraper to load the web page for a given amount of time before running the selector.

The Parent selector places the selector in the correct spot in the sitemap hierarchy, telling Webscraper.io that the current selector is extracting information from the homepage or on one of the pages to which another selector is linked. The Parent selector divides by hierarchy, as seen in the selector graph, with the homepage listed first as '_root,' followed by the selectors on that page. Those further down the list are selectors that usually link from the homepage to other pages in the website.

Figure 15—'Parent Selectors'

Ensure that everything is correct before saving your selector so that the wrong sitemap is not selected, or that no sitemap is selected at all. If the user does not highlight a Parent selector before saving it, the selector disappears altogether even though Webscraper.io thinks it is there. This means that the ID is in use but does not appear anywhere in the sitemap or in the exports. There is no solution for those selectors that cannot be found, which is bug in the software of which Webscraper.io is aware. A mislabeled selector will still appear, but it must be found manually and then edited to point to the correct Parent selector. The best way to find this type of error is to use the Selector Graph to show all selectors, minus the instance just discussed. Once found, users can navigate to the Parent selector in the selector menu and change it accordingly.

5. Scraping a Website

Scraping a website has three main steps:

- Creating a sitemap and selectors
- Extracting information
- Exporting information

This section discusses the second step. (See Sections 1 to 4 for step one and Sections 6 to 8 for step three.) Everything up to this point set up Webscraper.io to perform a scrape of a website. Creating a sitemap and selectors tells the system what to do during the scraping process. Then, users tell Webscraper.io to go through all selectors and perform the actions set within them with the Scrape panel, when the data is actually extracted from the website. The scraper uses the information pulled here to generate previews and export files.

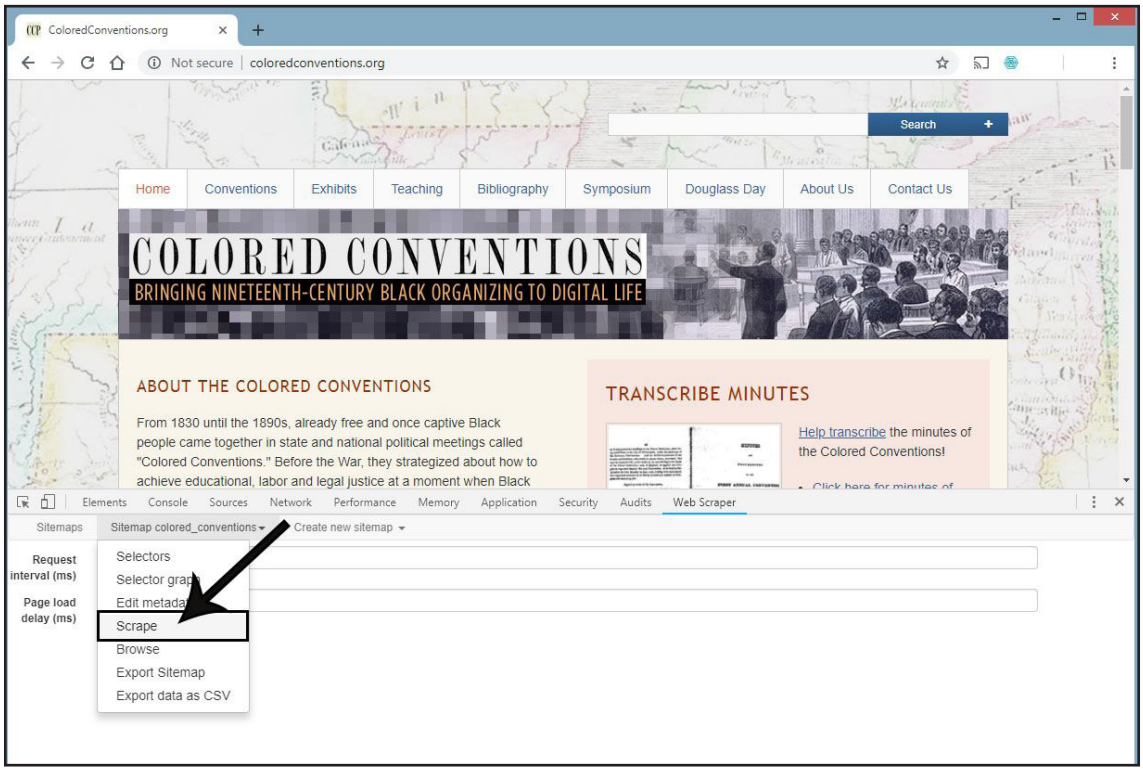


Figure 16—'Scrape' button

Users have the option to add either a request interval or a page load delay to the entire scraping process. (See Creating a Selector). With

both options, the scraper loads pages with different timing so that websites can load information before the scraper begins extracting information. The time delay is in milliseconds, with a default of 2000. Anything shorter than this may mean that the page has not loaded information for scraping. Both options add time for a page to load in case there is a lot of information, or if there are elements that take more time to load. Once the preferred time is entered, click the 'Start scraping' button.

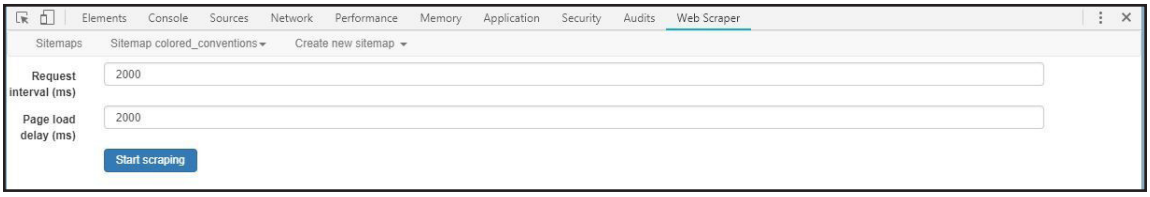


Figure 17—Scrape panel

Once started, a scrape opens a new browser window containing the web page to which a selector has been directed and cycles through it. There is no indication that the information is being extracted, but the windows will open and close as the scraper goes through the sitemap. When the scrape is finished, a pop-up window appears in the bottom right corner of the computer screen stating the scrape is done.

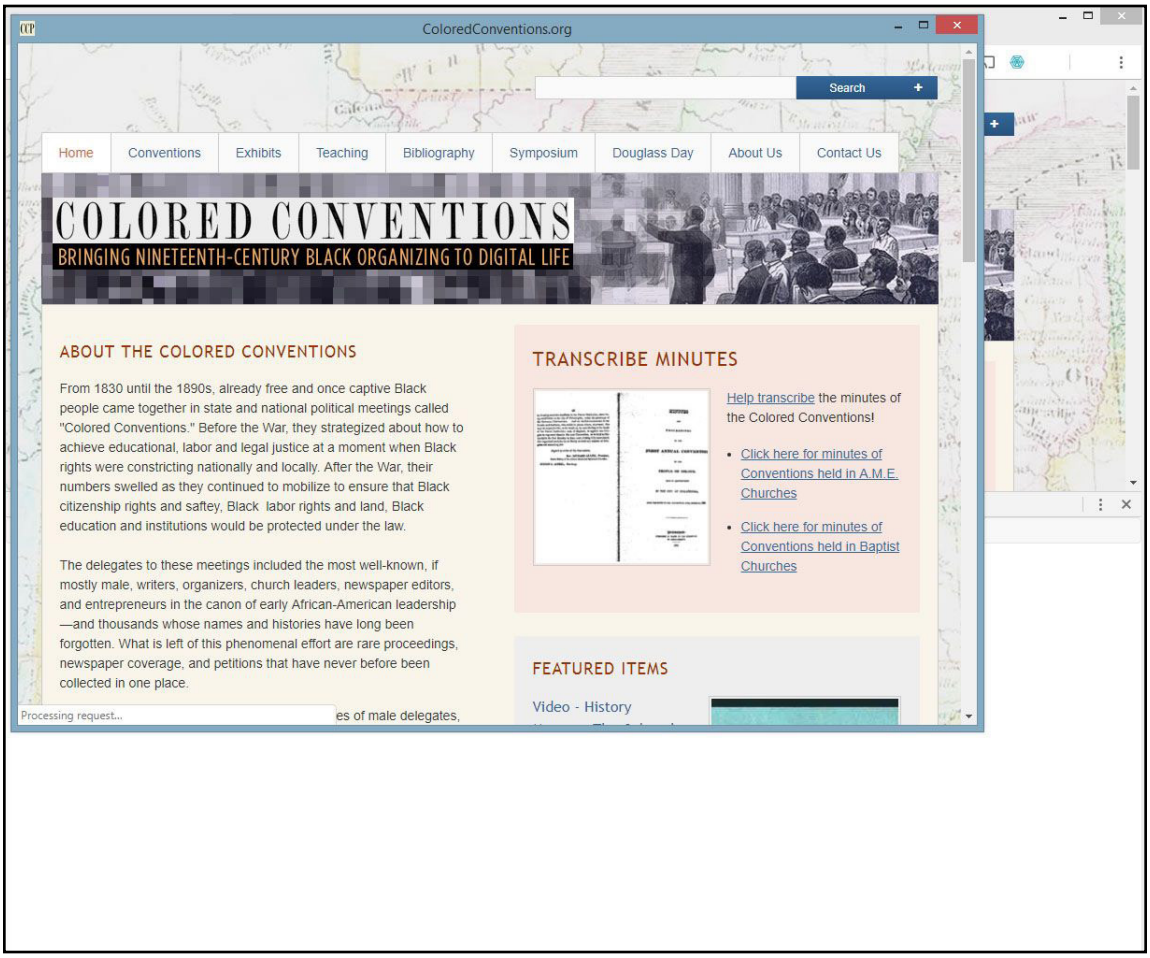


Figure 18—Scrape window

The web scraper automatically directs to the Browse panel when it is finished. Clicking on the 'Refresh' button shows the data preview.

Figure 19—Finished scrape

6. Browsing Scraped Data

Any scraped data can be viewed by accessing the Browse panel. The scraper also automatically redirects to this panel when a scrape is finished.

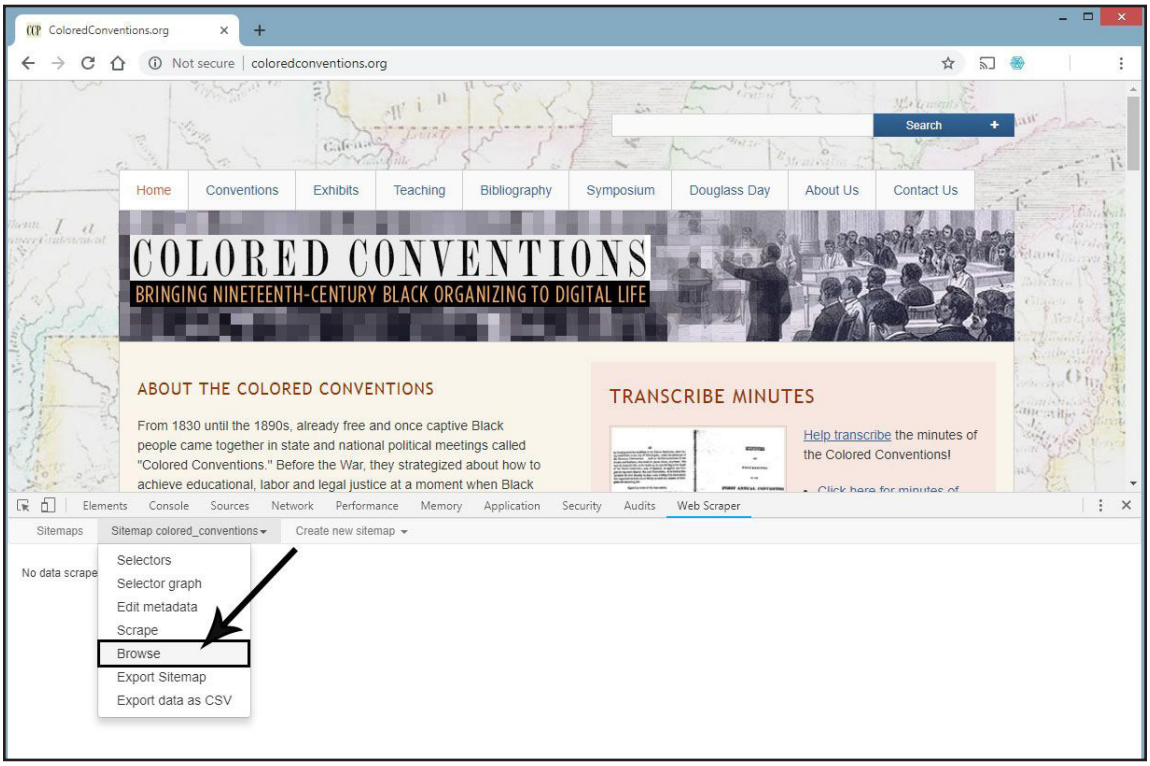


Figure 20—Browse data

If no data appears, click 'Refresh.'

Figure 21—Refresh data

Webscraper.io sets the data up as a spreadsheet and provides a preview of the data prior to downloading the CSV file (See Figure 22). This helps users ensure all the data is present and accounted for, including the information present in the various HTML and CSS tags from all the selectors in the sitemap. Note that the selector ID is now the column header. The exported CSV file will structure the data the same way as the preview.

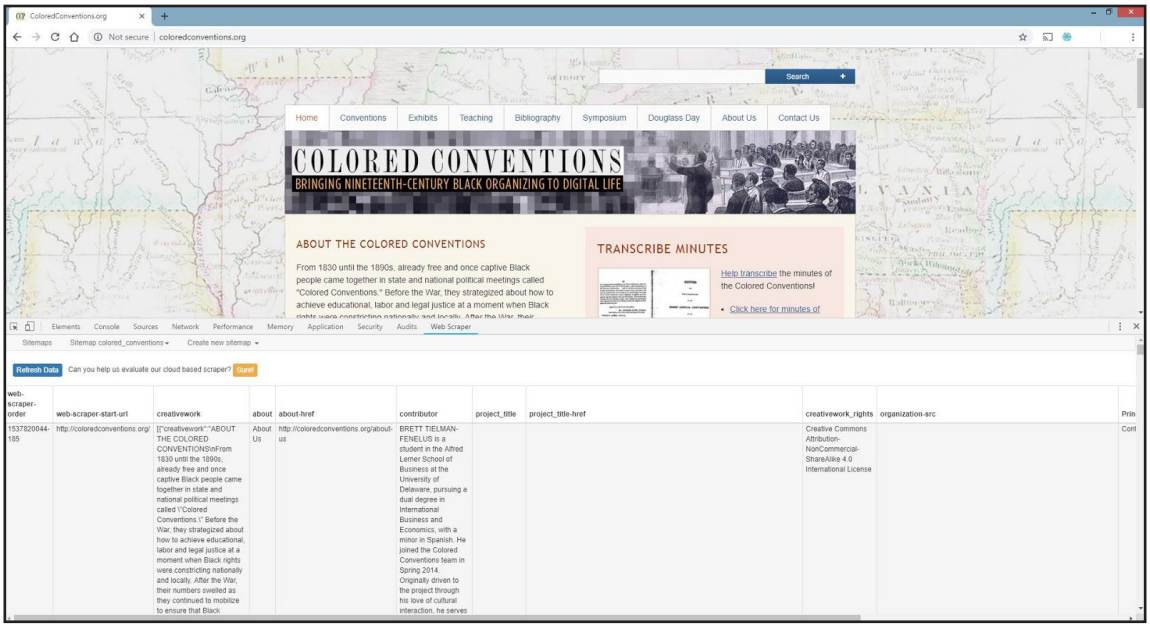


Figure 22—Preview of data

7. Exporting Sitemaps

Any sitemap with information can be exported using the 'Export sitemap' panel. Exporting a sitemap involves all the information except scraped data, such as the sitemap name, starting URL, and all the selectors created in the sitemap (see Exporting Data).

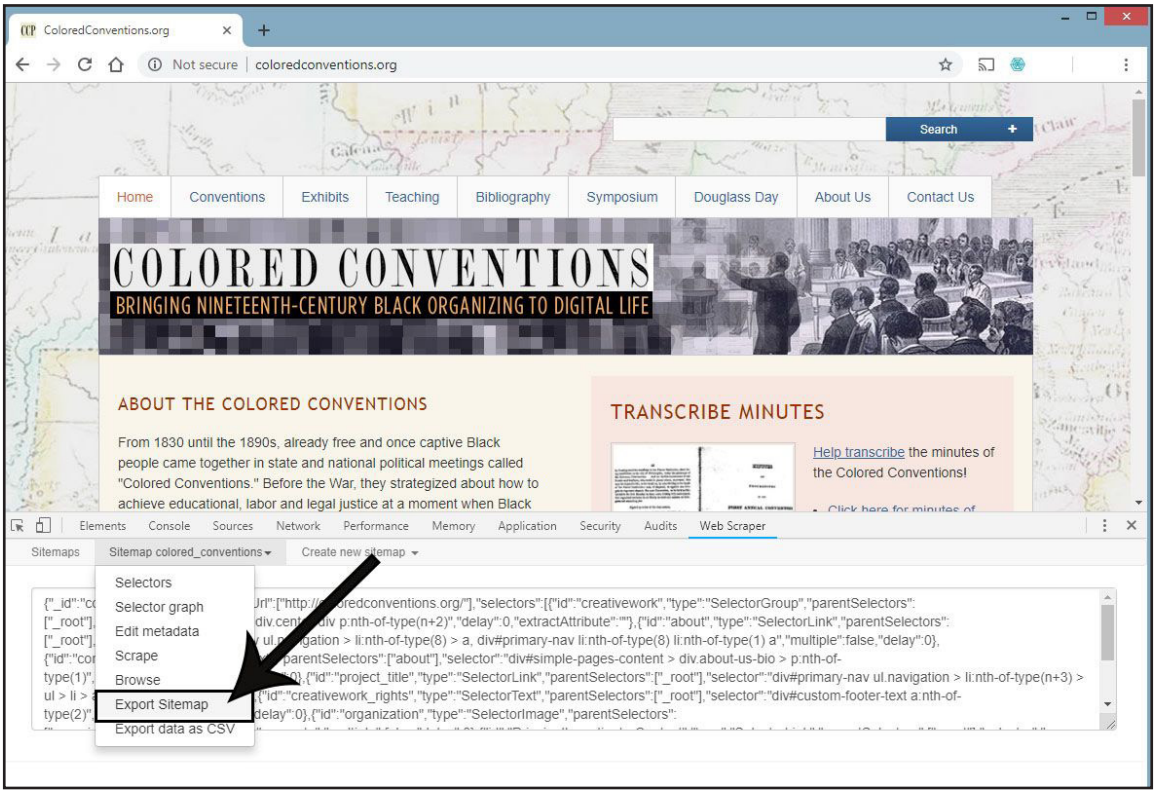


Figure 23—'Export Sitemap' button

The sitemap export generates JSON code in the box that opens when users access the panel. The safest way to copy the code is to click within the box and then hit CTRL+A to select all text. Users can then copy the code by either pressing CTRL+C or right-clicking their mouse and selecting 'Copy.' The code can then be pasted as a text file into a word processor to save a copy or into an email for sharing. Any changes elsewhere in Webscraper.io will also change this export, so any previously saved sitemaps will not be accurate.

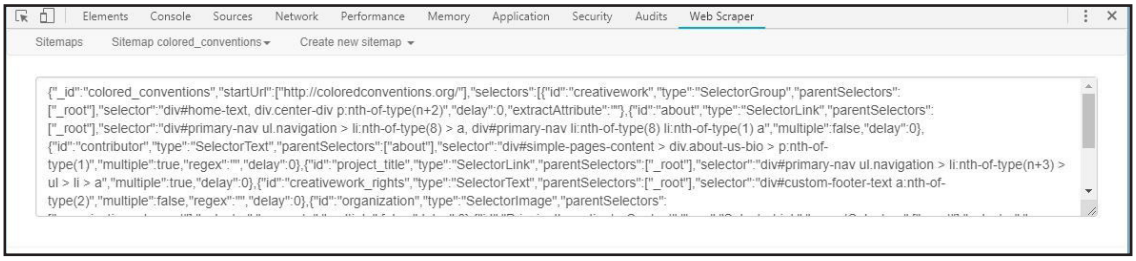


Figure 24—Export sitemap panel

8. Exporting Data

Webscraper.io exports the scraped data through the 'Export data as CSV' panel. This is different from exporting a sitemap (See Exporting Sitemaps), as this panel downloads a CSV file to the user's computer. Users must have already done a scrape of the website to extract the information (See Scraping a Website). This is typically the last step in a scraping project, as it is the final output of Webscraper.io.

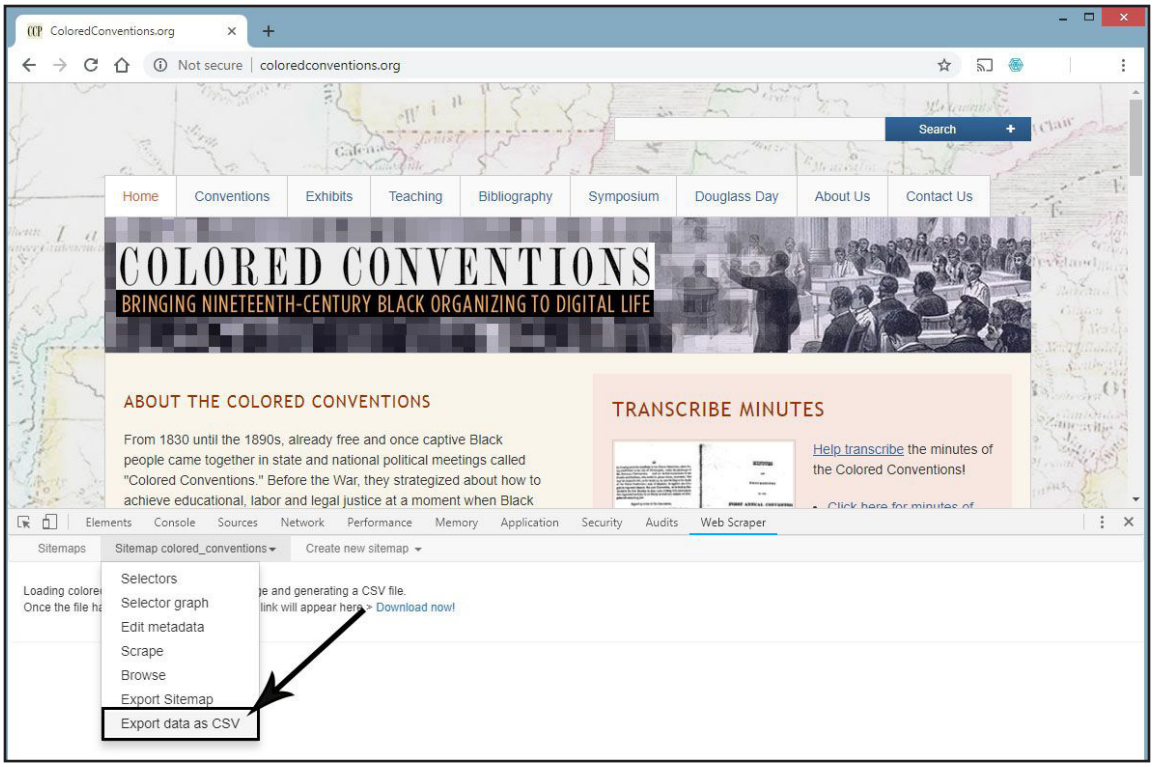


Figure 25—'Export data as CSV panel

A downloadable file generates as soon as a user enters this panel, with a blue 'Download now' link appearing when the file is ready for download. Once clicked, the file downloads to the location set in the browser settings. A pop-up box also appears at the bottom of the page, which users can use to open the file directly.

Figure 26—Export data as CSV

CSV files can be opened in spreadsheet software like Excel or Google Sheets, as opposed to a word processor. Users can also convert them to another file type fairly easily.

Appendix B

To view an example of a Webscraper.io CSV Export file, visit Florida Memory Export September 6, 2018

Endnotes

1. Michael Roth created this document in October 2018 as part of the SHARE Initiative. Contact him via email at: michael.roth89@gmail.com.
2. See http://www.share-research.org/.
3. See https://omeka.org/classic/directory/ for the list.
4. ParseHub is found at https://www.parsehub.com/.
5. The extension is found at webscraper.io.
6. https://www.floridamemory.com/
7. http://newdeal.oucreate.com/
8. http://omeka.uws.edu.au/farmstofreeways/
9. http://mallhistory.org/
10. http://coloredconventions.org/
11. https://www.floridamemory.com/
12. http://gothicpast.com/
13. https://share.osf.io/api/v2/schema/
14. Metadata Dictionary
15. See http://nthmaster.com/ for a detailed explanation.

ASSOCIATION OF RESEARCH LIBRARIES

Association of Research Libraries
21 Dupont Circle, NW
Suite 800
Washington, DC 20036
T 202.296.2296
F 202.872.0884

ARL.org
pubs@arl.org